2003

# Molecular phylogenetics and evolution in sharks and fishes

Juan Andrés López
*Iowa State University*

# Molecular phylogenetics and evolution in sharks and fishes

by

**Juan Andrés López**

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Ecology and Evolutionary Biology

Program of Study Committee:
Gavin J. P. Naylor, Major Professor
Drenna Dobbs
Frederick Janzen
Christopher Tuggle
Jonathan Wendel

Iowa State University

Ames, Iowa

2003

Graduate College

Iowa State University

This is to certify that the doctoral dissertation of

Juan Andrés López

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

To my wife Ellen, my parents Javier and Myriam, and my sister Emilia.

# TABLE OF CONTENTS

# CHAPTER 1. GENERAL INTRODUCTION

## Introduction

The component of evolutionary biology that focuses on the development and testing of hypotheses of the evolutionary relationships, or phylogenies of biological lineages has undergone important changes in the last four decades. The two chief causes of these changes were: (1) the development of a formal body of theory connecting the observed variation among species to the phylogenetic hypothesis implied therein, and (2) the continued development of technologies for the fast and easy determination of the composition of genes. The latter cause is also responsible for the democratization of this once rarefied field. And, with democratization has come a plurality of opinion regarding all aspects of phylogenetic inference.

As a consequence of the diversity of perspectives prevalent among researchers practicing phylogenetic research there are intense disagreements of opinion revolving around two aspects: (1) the ability of different methods of inference to capture the history of relationships among lineages; and (2) the relative merits of evidence derived from observations of the morphological and genetic variation among species and groups of species. In this context, this dissertation intends to explore ways to develop defensible phylogenetic hypothesis based on genetic evidence by the judicious application of the array of methods of inference currently available followed by a comparison of the information provided by genetic evidence to that supporting any previously proposed hypotheses. Thus, the primary goal of the case studies that compose this dissertation (Chapters 3 through 5) was to advance our understanding of the phylogeny of the groups of study only to the extent that the newly obtained molecular evidence permitted and to the extent it could be shown to

represent an improvement over existing hypotheses. This required close examination of both the molecular evidence as well as previously available relevant evidence to determine the quality of historical information preserved in the distribution of variation (molecular and morphological) among the surviving members of the three lineages examined.

**Thesis Organization**

This dissertation is composed of this introduction, four main chapters, each formatted as a journal article following the stylistic guidelines of journal appropriate for their content, and a general conclusions chapter. Of the four main chapters, only Chapter 2 is of my sole authorship. On the remaining three I share authorship with colleagues from the University of Nebraska and with members of the laboratory headed by my major professor, Dr, Gavin Naylor. In the three chapters where I share authorship, I have been responsible with conducting varying proportions of the data collection efforts, and all of the analyses, interpretation, synthesis and reporting efforts. Following is a broad overview of the subjects treated in each of the chapters and an explanation of how the work described in those chapters relates to the questions of phylogenetic inference that are the focus of this dissertation.

The first of the four main chapters of this dissertation is titled "Molecular Evolution of the *recombination-activating gene* (RAG1)." It is an examination of different aspects of the evolutionary properties of the RAG1 gene at the level of its coding DNA sequence and of its amino acid sequence. I compiled representative sequences of the RAG1 gene from a diverse array of vertebrate species using repositories of publicly available genetic data and newly acquired data. The chapter describes the results of a series of analyses aimed at understanding the various forces shaping the evolution of this gene as it relates to its value as

a source of phylogenetic information. For this reason the study focuses on aspects of molecular evolution such as base and amino acid composition, codon usage bias, and heterogeneity of rates of evolution which are known to present problems to the retrieval of phylogenetic information from patterns of variation in genetic sequences. The information derived from the analyses presented in this chapter was used in the interpretation of the results obtained in the following three chapters in which this gene, in combination with other molecular data, is used as a source of evidence to develop and test phylogenetic hypothesis in three different vertebrate lineages.

Chapter 3, entitled "Esociform relationships", is the result of collaboration with researchers at the University of Nebraska. The study described in this chapter is concerned with the phylogenetic relationships of a relatively minor – in terms of number of species and importance to humans – groups of fishes, the Esociformes. The esociform fishes, commonly known as pikes and mudminnows, have proven problematic to researchers attempting to determine their placement among other fish lineages of similar age. Using evidence from genetic sequences obtained from a sample that includes representatives from a broad diversity of fish species selected for their potential relationship to esociforms, we show that there is a strong case to reconsider currently accepted views on esociform relationships. In addition to the genetic evidence, we show that a close examination of the morphological evidence that supports current hypotheses of relationships of these fishes suffers from some critical flaws. Thus, strengthening the case for a revised esociform classification. This study highlights some of the issues that arise when molecular evidence supports a conclusion that contradicts hypothesis based on morphological data. The solution in this case proved to be a re-examination of the existing body of morphological evidence.

Chapter 4, entitled "Much More Data and Exhaustive Species Sampling does Not Improve Resolution of the Phylogeny of Lamniform Sharks", is the result of research conducted in cooperation with Dr. Gavin Naylor and two other students in his group. Aside from advancing our understanding of the evolutionary history of a group of sharks that includes several species with unusual biological characteristics (planktivory and endothermy), an important objective of this study was to explore the limits of phylogenetic inference given the nature and dynamics of evolution that are unique to a given biological lineage. In the field of phylogenetic inference, it is expected that increases in the amount of data brought to bear on a phylogenetic question and/or increases in the number of taxa included in a study will result in a phylogenetic reconstructions of increased accuracy. The study described in this chapter tests that expectation by examining a substantial amount of molecular data from all known living representatives of a group of sharks whose relationships had proven problematic in prior phylogenetic treatments. Contrary to expectations, the larger and more complete data set did not throw new light on the relationships among species in this group of sharks. We conclude that while there are reasons to be optimistic about the information that a larger data set may provide, it is important to be cautious to prevent the incorporation of method driven conclusions into working hypothesis of relationships. In cases similar to the one treated in this chapter, it may be preferable to adopt a conservative stance and accept a hypothesis with low information content (i.e. tree topologies with low resolution). The rationale behind this approach is that a poorly resolved hypothesis is more useful than an inaccurately resolved one.

Chapter 5, entitled "Phylogenetic Relationships Of Sharks Of The Family Triakidae (Carcharhiniformes: Elasmobranchii)", also resulted from collaborative work with members of Dr. Naylor's laboratory. In contrast to the other phylogenetic studies presented in this

dissertation where previous workers had advanced relevant phylogenetic hypotheses, the work described in this chapter represents the first attempt to elucidate evolutionary relationships among members of this commercially important group of sharks using molecular sequences as sources of evidence. Prior to this research the bulk of our understanding of the evolutionary history of this group was the result of the work of a single expert, whose output, while generally accepted, is difficult to review critically because the conclusions presented in those works are not always explicitly justified. This situation made the interpretation of the molecular-based results in this chapter different from that presented in the other studies because there were no previously accepted hypothesis to compare with the one supported by the molecular data. For this reason we opted to be maximally conservative and only accept those results that were unambiguously supported by all the analyses. Another important characteristic of this study is that the number and relative abundance of species in this group of sharks made it impossible to assemble samples from all members of the group. Incomplete sampling is a common in phylogenetic studies but, unfortunately, the conclusions reported in many of the studies with incomplete sampling fail to draw attention to the limitations placed by the absence of certain taxa in the analysis. The phylogenetic conclusions advanced in this chapter take into account the absence of certain members of the group from the taxonomic sample used in the study.

Chapter 6 is a summary of the main themes explored in the four preceding chapters. This chapter also outlines some general conclusions regarding the theory and practice of phylogenetic inference as applied in this dissertation and in current research in the field. In addition, the chapter outlines some potential areas of future research that would further our understanding of the limits of phylogenetic inference and, therefore, help to better determine the confidence that competing hypothesis of phylogeny should be afforded.

# CHAPTER 2. MOLECULAR EVOLUTION OF THE *RECOMBINATION ACTIVATING GENE* 1 (RAG1)

A paper to be submitted to *The Journal of Molecular Evolution*

Juan Andrés López

## Abstract

DNA sequences from the recombination-activating gene 1 (RAG1) are being used in a growing number of studies of the phylogenetic relationships of different vertebrate lineages. I have assembled a diverse set of RAG1 sequences from four vertebrate clades and analyzed their evolutionary characteristics to better understand the phylogenetic value of this gene and to delineate regions of the protein that may be functionally important. This analysis showed that RAG1 sequences are subject to variable evolutionary constraints between different sites on the molecule and between different lineages. As a result, substitution characteristics and the rate of evolution of the DNA and amino acid sequences are heterogeneous between and within the vertebrate clades considered here. This heterogeneity must be given consideration when devising phylogenetic analyses based on RAG1 sequences. The comparative sequence analysis highlighted two highly conserved regions with no known functions. The first of these is located in the N-terminal end of the molecule and is restricted to bony vertebrates. The second is conserved in all known sequences and contains three pairs of closely spaced cysteine residues. In addition, the fish sequences examined show some unique features not observed in other vertebrates. The regions delineated by sequence comparison should help guide studies of RAG1 function, which is critical in the development of the adaptive immune system through its mediation of the mechanism of antigen receptor loci rearrangement known as V(D)J recombination.

## Introduction

DNA sequences from the recombination-activating gene 1 (RAG1) are gaining popularity as a source of phylogenetic information in diverse vertebrate lineages (e.g. passerine birds, Barker et al., 2000; placental mammals, Waddell and Shelley, 2003; elasmobranchs, G. Naylor, pers. comm.; teleost fishes, López et al., accepted). As the number of known RAG1 sequences and the taxonomic diversity they represent grow it becomes useful to study patterns of conservation and change along the length of the molecule with two aims in mind. First, information derived from such an undertaking may guide investigations into the functional constraints of the gene product, particularly in parts of the molecule that differ between major clades or that have not been amenable to experimental manipulation. Second, an examination of the DNA sequences using current knowledge of the functional constraints of the protein may be informative in the selection of the methods used to recover phylogenetic information from them.

The goal of the present study is to describe and contrast patterns of DNA and amino acid sequence evolution along the RAG1 coding sequence from representative shark, fish, bird and mammalian taxa in the context of current knowledge of RAG1 function. A secondary goal is to determine whether or not the evolutionary characteristics of the gene can be expected to make its DNA sequence or subsets of it suitable for phylogenetic studies.

### *RAG1 and the adaptive immune system*

The vast diversity of antigen receptor specificities that characterizes the adaptive immune system is generated at the genetic level by a tightly regulated recombination process that results in the rearrangement of the immunoglobulin (Ig) and T-cell receptor (TCR) loci and the generation of novel Ig and TCR gene variants. This process is termed V(D)J

recombination in reference to the variable, diversity and joining gene segments that are brought together to form a novel part of the functional gene. The diversity segments are part of only three (IgH, TCRβ and TCRδ) of the seven antigen receptor loci in mammalian genomes.

Prior to rearrangement, an Ig or TCR locus may typically contain tens to hundreds of segments with the potential to generate up to $10^4$ combinations. In addition, the imperfect joining of the segments during V(D)J recombination has been estimated to increase the number of potential protein variants to $10^{12}$ (Davis, 1990). In Ig loci, this antigen receptor diversity is further modified and adapted to specific antigen challenges through the combined action of 'class or isotype switching' and 'somatic hypermutation' (Sadofsky, 2001).

The number of factors known to mediate V(D)J recombination *in vivo* is still growing (e.g. Moshous et al., 2001; Dai et al., 2003) and there are important aspects of the process, such as gene segment selection, that remain poorly understood (Livak and Petrie, 2002). Most of the factors implicated in V(D)J recombination are not specific to lymphoid cells and have other roles in non-homologous end joining (NHEJ), DNA break repair and DNA remodeling (e.g. DNA-PKcs, Ligase IV, HMG-1). Two important exceptions are the lymphoid specific proteins RAG1 and RAG2, which have been shown capable of catalyzing the initial stages of V(D)J recombination *in vitro* in the absence of any other factors, albeit with lower efficiency than in their presence and, which have no other known functions *in vivo* (but see Agrawal et al., 1998 and below). At present, the RAG1/2 recombinase is the only known V(D)J recombination-specific factor.

*V(D)J recombination*

The process of V(D)J recombination begins with the recognition of recombination signal sequences (RSS's) that abut the antigen receptor gene segments. An RSS consists of a highly conserved heptamer sequence flanking the coding segment (CACAGTG, where the three sites closest to the coding sequence exhibit the greatest degree of conservation) and a non-conserved spacer sequence of 12 or 23 base pairs separating the heptamer from a moderately conserved AT-rich nonamer sequence (Fig 2.1A). Nonamer binding by the RAG1/2 recombinase precedes heptamer recognition, but while nonamer absence only reduces the efficiency of the reaction, the lack of the heptamer eliminates it (Godderz et al., 2003). At each antigen receptor locus, the distribution of 12 and 23bp RSS's distinguish the different gene segment types (Fig 2.1B). To ensure that the right combination of gene segments are joined, recombination between segments with different spacer lengths occurs with much greater efficiency than between segments of similar spacer lengths. In addition, CpG methylation, chromatin organization and accessibility may also have a role in determining which segments are available during V(D)J processing (Roth and Roth, 2000; Nakase et al., 2003).

After signal recognition and binding, the RAG1/2 recombinase introduces a nick between the 5' end of the heptamer on the RSS and the coding segment. The freed hydroxyl group on the 3' end of the coding segment and the opposite strand join via transesterification generating a hairpin on the coding end and a blunt end on the signal segment. The hairpin coding ends are subsequently nicked, most frequently in an asymmetrical manner that results in an overhang of a few nucleotides. The overhangs are then filled to produce blunt ends with added base pairs relative to the original coding end (Fig 2.1C). Finally, two blunt, modified coding ends are joined to produce a coding joint between two different antigen

receptor gene segments. The signal end is not modified and is subsequently joined to another signal end, thus preventing potentially harmful translocations. Recombinant hybrid joints between signal and coding ends are rare at antigen receptor loci but are observed frequently in RAG1/2 recombination assays indicating a role of other factors in regulating proper V(D)J recombination.

*RAG1 structure and function*

Because the RAG1 protein, in conjunction with RAG2, mediates key steps in the V(D)J recombination reaction, it has been the focus of numerous studies aimed at determining its catalytically active residues and structurally important regions and their mode of action. The RAG1 protein must play a number of different roles to mediate a successful V(D)J recombination reaction. Accordingly, mounting evidence shows RAG1 to be a complex molecule consisting of structurally and/or functionally distinct domains.

The functions of RAG1 may be artificially classified into two general categories (Table 2.1). First, functions that pertain to interactions with other factors required for V(D)J recombination. Functions that may be grouped in this category are sub-cellular localization, oligomerization, recombinase formation with RAG2 and the as yet undetermined interactions with the other factors that are involved in V(D)J recombination. Second, intrinsic RAG1 functions required in the recombination reaction proper. These include the recognition of the RSS's and the hydrolysis of the DNA backbone at the junction between coding and signal sequences. It is unclear at this point what roles the RAG1/2 recombinase has after DNA cleavage, but any such role for RAG1 would fall under this category. The results of mutational studies have delineated many of the regions of the RAG1 molecule that are required to carry out these functions (outlined in Table 2.1 and summarized below).

A truncated RAG1 protein comprising amino acids 384-1008 of the *Mus musculus* RAG1 protein can mediate V(D)J recombination in conjunction with RAG2 on exogenous substrates *in vivo* (Sadofsky et al., 1993). This segment is often termed 'core RAG1' and has been used in many experimental studies of RAG1 function because it is more amenable to manipulation than the full-length protein, which is poorly soluble. Although the so-called dispensable regions that fall outside of core RAG1 exhibit less sequence conservation compared to the core protein, there are stretches within these regions with recognizable homology across all known sequences. In fact, the N-terminal portion of RAG1 outside of the core region has important functions in antigen receptor locus development. For example, recombination mediated by core RAG1 is significantly less efficient than that mediated by the full-length protein due in part to the exclusion of a stretch rich in basic amino acids between residues 216 and 238 (McMahan et al., 1997; Steen et al., 1999). Also, rearrangement of Ig, but not TCR, loci was compromised in a patient homozygous for a deletion in RAG1 that leads to the synthesis of an N-terminal truncated protein (residues 203-1040 of the murine sequence) through the use of an alternative start codon (Noordzij, 2000). And, truncated RAG1 produces more hybrid joints (signal to coding segment) than the full-length protein in assays *in vivo* (Sekiguchi et al., 2001).

V(D)J recombination takes place in the nucleus, therefore upon synthesis of the recombination factors must be directed to that cellular compartment. The dominant nuclear localization signal in RAG1 is located in the C-terminal region along two segments rich in basic amino acids. Both of these segments, spanning residues 826-840 and 969-973 of the murine sequence, bind importin $\alpha$ 2 subunit and when deleted or mutated disrupt RAG1 localization to the nucleus (Cuomo et al., 1994; Spanopoulou et. al., 1995). In addition, three other segments rich in basic amino acids located between positions 216 and 270 interact with

importin α 1 subunit; however deletion or mutation of these basic motifs does not completely disrupt importin binding or sub-cellular localization making the functional relevance of these sequences uncertain (Cortes et al., 1994; Spanopoulou et. al., 1995). This five basic amino acid motifs have been termed BI through BV in the order of their location on the protein in the N-terminal to C-terminal direction. Importins recognize nuclear localization signals and form complexes that interact with nuclear pore proteins to mediate nuclear transport.

There is uncertainty regarding the stoichiometry of the RAG proteins in the active recombinase. Some evidence suggests that RAG1 forms stable dimers and combined with one or two RAG2 molecules make a functional recombinase (Swanson, 2002), while other results suggest that RAG1 forms tetramers in the functional recombinase (Godderz et al., 2003). Several regions of the RAG1 protein have been implicated in oligomer formation (Rodgers et al., 1996; Arbuckle et al., 2001). A distinct structural domain located N-terminal to core RAG1 mediates stable dimerization (Rodgers et al., 1996). This domain is composed of a RING and a $C_2H_2$ zinc finger subdomains. The RING finger extends between residues 286 and 330 and binds two zinc ions on its own and a third one in conjunction with residues 266 and 270. The Zinc finger (termed ZFA) extends between residues 350 and 378 and binds one zinc ion. The tertiary structure that this region of the molecule acquires upon binding the zinc ions propitiates the presentation of two helices that form the dimerization interface. On the dimerization interface, three phenylalanine (Phe) residues on each member of the dimer are proposed as important hydrophobic factors in the stability of the dimer (Bellon et al, 1997). Interestingly, recent evidence suggests that the RING finger of RAG1 may also function as an E3 ubiquitin ligase (Yurchenko et al., 2002). The dimerization domain lies outside of core RAG1, which forms stable dimers on its own (Rodgers et al., 1999). To date, the core RAG1 regions implicated in self-association are the C-terminal

domain (Arbuckle et al., 2001; see below) and an N-terminal region (unpublished data of P. De and K. K. Rodgers cited by Godderz et al., 2003).

The interaction between RAG1 and RAG2 is essential for V(D)J recombination. At least in part, this interaction has been mapped to the second zinc finger (or ZFB) of RAG1, which spans the interval 727-750 (Rodgers et al., 1996), and to the sixth repeat of a series of six putative kelch motifs that constitute the functional core of RAG2 (Aidinis et al., 2000). ZFB also plays a role in binding the heptamer sequence of RSS's (Peak et al., 2003). Little is known about specific interactions between RAG1 and the non-lymphoid specific V(D)J factors, such as HMG1 and DNA-Pkcs.

Aside from interacting with other recombination factors, RAG1 has important roles in RSS's binding and DNA cleavage. Core RAG1 includes two distinct topological domains (central and C-terminal) and an N-terminal region; each with distinct DNA binding properties (Arbuckle et al., 2001). The nonamer-binding region (NBR, residues 389-445), which shows specific affinity for the nonamer sequence of RSS's, lies within the N-terminal region (384-527) of core RAG1 (Spanopoulou et al., 1996). The NBR and adjacent sequences show significant sequence homology to the bacterial Hin recombinases and the nonamer sequence is similar to the Hin recognition sequence. These observations have invited speculation about the origin of the RAG1 gene (Bernstein et al., 1996; Banerjee-Basu and Baxevanis, 2002 and see below).

The central domain of core RAG1 (528-760) binds with high specificity to the heptamer sequence of RSS's. The affinity of this interaction is stronger when the heptamer DNA is single stranded (Peak et al., 2003). The C-terminal domain (761-979) contains a region (889-974) that makes close contact with the coding segment during recombination (Mo et al., 2001). Binding of an RSS by the RAG1 dimer results in radical changes in the

topology of the latter. These changes include the exposure of buried residues to the solvent surface (Ciubotaru et al., 2003).

A single RAG1 molecule contains a functional active site for the nicking and hairpin formation reactions (Landree et al., 1999). To date, five amino acid residues (D600, D708, R713, H795 and E962) are known to be essential for these steps of V(D)J recombination (Landree et al., 1999; Huye et al., 2002). In addition, mutations studies have shown that the recombination efficiency of core RAG1 is very sensitive to deletions and mutations probably as result of the many different interactions mediated by this protein (Landree et al., 1999; Huye et al., 2002).

*RAG1 origin and phylogenetic distribution*

One of the most interesting aspects of the evolution of the adaptive immune system is its origin. The adaptive immune response has been documented among all groups of jawed vertebrates, albeit with differences in specific aspects of the phenotype (e.g. the diversity of antigen receptors is much reduced in sharks). Aside from antigen receptor loci, the other major genetic components of the adaptive immune system are the MHC class I and class II loci. All of the major components have been found in chondrichthyans (sharks and their allies), the most ancient lineage known to exhibit an adaptive immune response, but have not been found in older lineages. Similarly, the RAG1 and RAG2 genes are present in chondrichthyans, but all attempts to ascertain their presence in the genomes of more ancient lineages have failed. Further, efforts to determine the homology of the RAG proteins have failed to find likely relatives among eukaryotic genomes (Banerjee-Basu and Baxevanis, 2002).

In all cases documented to date the RAG genes reside in close proximity in the genome and are oriented tail to tail (Schluter and Marchalonis, 2003). Also noteworthy, is the fact that the RAG1/2 recombinase can catalyze two types of transposition reaction *in vitro* (Agrawal et al., 1998; Shih et al., 2002). And, the mechanism of V(D)J recombination bears similarity to the transposition mediated by bacterial transposases and retroviral integrases (van Gent et al., 1996). Overall, the evidence suggests that the RAG genes invaded the genome of an ancestor of the jawed vertebrates by horizontal transfer and may be related to transposable elements.

While mutational studies are required to establish with certainty the contribution of different residues to the function of a protein, comparative studies of gene sequences from different lineages can complement that knowledge by revealing patterns of variation such as the distribution and extent of conserved regions along the molecule and their relationship to functionally important regions or residues. The present study highlights patterns of conservation in poorly understood regions of the molecule and compares characteristics of RAG1 evolution in representatives of four vertebrate groups. These observations may help guide future studies on RAG1 function and inform the selection of methods in phylogenetic studies based on this gene.

## Materials and Methods

### *DNA sequences*

RAG1 coding sequences from representatives of four important vertebrate clades were aligned and compared. These four groups are: carcharhiniform sharks, euteleostean fishes, passeriform birds and therian mammals. The sequences used to represent the last two groups were obtained from Genbank. The passeriform bird sequences were originally

determined and reported by Barker et al. (2002). Fourteen sequences were selected to broadly sample the range of divergence within the passeriform order. In addition, the sequence from the galliform *Gallus gallus* was included to increase the level of divergence shown by the bird sequences. All seven full-length mammalian RAG1 sequences available in Genbank were used to represent that group.

The carcharhiniform RAG1 sequences were determined by G. Naylor and co-workers as part of a large scale phylogenetic study of galeomorph sharks. These sequences were determined by sequencing of PCR products generated from genomic DNA templates. Targeted PCR amplification and sequencing employed a battery of primers designed with specificity to shark species (G. Naylor, pers. comm.). The carcharhiniform sequences examined in this study were selected to broadly sample the range of divergence within the group. Three complete RAG1 sequences from euteleost species from Genbank and newly reported sequences from seven species of the euteleost order Esociformes were used to represent the fish clade. The esociform RAG1 sequences were determined by PCR and PCR-product sequencing with the following set of primers (relative primer locations given in Fig 2.2): 1AF-aagctgttccgggtcaggtcdatgga; 1AR-gcatcatggcctccagttcgtcngc; 1BF-agagtggacctgaarctscaggag; 1BR-cttgcatcatggcytccagttc; 1CF-ctgcaaatccaccagctggcca; 1CR-gcagctgcggcagaagaggtgt; 1DF-gtcacytgtcaggtstgtgaycacct; 1DR-bkwyttggtgtctccaccttcctc; I2AF-gaagacccaagtaaaggtgtttgc; I2AR-gtggtagcctggtagtagctcctt; I2BF-gaactggargccatgatgcaaggt; I2BR-tggctrcagctcaggaaygtgttgac; 2AF-ctgagctgcagtcagtaccataagatgt; 2AR-ctgagtccttgtgagcttccatraaytt; 2BF-aggagtcctgcgatggyatggg; 2BR-cctctgarcrcaccagctcaca; 2CF-ttygtggaygagtcrgaccacgagac; 2CR-gtagttkgtgatcttgccgtcrtacc.

Following are the Genbank accession numbers of the sequences not reported here as new, and the abbreviation used to denote each sequence in all figures and tables in this article. From Carcharhiniformes: *Carcharhinus leucas* (U62645) Carleu; from Euteleostei: *Brachydanio rerio* (U71093) Brarer; *Takifugu rubripes* (AF108420) Takrub; and *Oncorhynchus mykiss* (U15663) Oncmyk; from Neognathae (Passeriformes and Galliformes): *Smithornis rufolateralis* (AY057031) Smiruf; *Tyrannus tyrannus* (AF143739) Tyrtyr; *Formicarius colma* (AY056993) Forcol; *Pardalotus striatus* (AY057015) Parstr; *Oriolus larvatus* (AY057011) Orilar; *Coracina lineata* (AY056988) Corlin; *Picathartes gymnocephalus* (AY057019) Picgym; *Sturnus vulgaris* (AY057032) Stuvul; *Certhia familiaris* (AY056983) Cerfam; *Aegithalos iouschensis* (AY056976) Aegiou; *Hirundo pyrrhonota* (AY056997) Hirpyr; *Oedistoma iliolophum* (AY057010) Oedili; *Prunella collaris* (AY057024) Prucol; *Emberiza schoeniclus* (AY056992) Embsch; *Gallus gallus* (M58530) Gallga and from Mammalia: *Homo sapiens* (NM_000448) Homsap; *Mus musculus* (AY215075) Musmus; *Rattus norvegicus* (XM_230375) Ratnor; *Oryctolagus caniculus* (M77666) Orycan; *Sus scrofa* (AB091392) Susscr; *Lama glama* (AF305953) Lamgla; *Monodelphis domestica* (U51897) Mondom. In addition, two full-length sequences from amphibian species were included in the analysis of sequence conservation. These species, accession numbers and abbreviations are *Xenopus laevis* (L19324) Xenlae and *Pleurodeles walt* (AJ010258) Plewal.

Following are the names and abbreviations of carcharhiniform and euteleost sequences from which newly reported RAG1 sequences are examined here. The abbreviations are used in all figures and tables of this article. From Carcharhiniformes: *Apristurus labiosus*, Aprlab; *Carcharhinus dussumieri* Cardus; *Chaenogaleus macrostoma* Chamac; *Eridacnis sp.* Eridac; *Galeus boardmani* Galboa; *Galeorhinus galeus* Galgal;

*Halaelurus lineatus* Hallin; *Leptocharias smithii* Lepsmi; *Mustelus henlei* Mushen; *Negaprion acutus* Negacu; *Poroderma pantherium* Porpan; *Rhizoprionodon acutus* Rhiacu; *Sphyrna zygaena* Sphzyg. From Esociformes: *Dallia pectoralis* Dalpec; *Esox americanus* Esoame; *E. lucius* Esoluc; *E. masquinongy* Esomas; *E. reichertii* Esorei; *Novumbra hubbsi* Novhub; *Umbra krameri* Umbkra; *U. limi* Umblim.

The DNA sequences were aligned with the computer program Clustal X (Thompson et al., 1997) using default settings. The automated alignment was then manually edited to adjust the distribution of gaps (i.e. to preserve reading frame) and to correct obvious flaws in the automated alignment. These flaws were detected with the guidance of the amino acid translation of the DNA sequences.

*Sequence analysis*

The ranges of uncorrected sequence divergence values within each clade were determined for different partitions of the data (e.g. N-terminal dispensable region, core region, codon positions) based on nucleotides and their amino acid translation. Likelihood ratio tests of molecular clock were performed for each clade for both the entire sequence and for the core region alone.

The base composition of each sequence at all sites and at each codon position was determined and tested for homogeneity within each of the four vertebrate clades using the $\chi^2$ test implemented in PAUP* 4.0b10 (Swofford, 1998). The base composition characteristics of intron and exon sequences were compared in the euteleost sequences. Base composition differences between species in a group and between groups were visualized using plots of nucleotide proportion. Codon usage tables, effective number of codons (Nc, Wright, 1990) and the proportion of guanines and cytosines at synonymous 3$^{rd}$ codon position sites (GC3s)

were calculated to determine the relationship between base composition biases in the DNA sequences and biases in synonymous codon usage. These indices were calculated with the computer programs GCUA (McInerney, 1998) and CodonW (J. Peden, unpublished); the latter as implemented in the Pasteur Institute web server (http://bioweb.pasteur.fr).

Patterns of change and conservation along the molecule were determined. These measures of variability were compared against the functional constraints known to affect different regions of the molecule. Special attention was given to non-conservative amino acid substitutions at critical sites and to conserved substitutions that defined the different clades. The computer programs AMAS (Livingstone and Barton, 1993) and ALSCRIPT (Barton, 1993 were used to facilitate these observations.

## Results and Discussion

### *RAG1 sequence divergence and lineage age*

As one of the goals of this study is to describe patterns of amino acid residue conservation to determine functionally constrained regions of the RAG1 protein, it is critical to first understand the different levels of divergence represented by the sampled sequences. This is important because the significance of conservation is only relative to the length of substitution time over which the site has remained unchanged. Substitution time can be understood as the amount of change that could have taken place given the rate of substitution and the time of origin of a lineage. For example, it has been observed that rates of molecular substitution are significantly lower in sharks than in other vertebrate groups (Martin et al., 1992), hence even as the time since the split of the shark lineage and the remaining vertebrates (the Teleostomi) is one and the same, the genomes of members of the shark lineage have accessed a smaller volume of the genotype space. That is, the age of the shark

and teleostom lineages are the same but the substitution time they have spanned is different. I measured pairwise sequence differences within each of the four vertebrate clades examined in this study and compared these values between clades to have an estimate of the substitution time spanned by each clade and better understand the significance of the patterns of variability observed.

Despite ongoing debate on the timing of the origin of modern orders of birds it is likely that the passeriforms and *Gallus gallus* represent the youngest divergence among the four groups of sequences examined here. The paleontological evidence suggests that modern birds radiated in the early Tertiary (60 mya; Feduccia, 1996), while estimates based on molecular data push that date back to the middle of the Cretaceous (90 mya; Hedges et al, 1996). The origin of the other three vertebrate clades in this study can be traced to the period surrounding the Jurassic-Cretaceous boundary. Fossil evidence of carcharhiniform sharks first appears in deposits from the Early Cretaceous (Shirai, 1996). Similarly, fossils ascribed to euteleostean species are as old as the Early Cretaceous (Colbert et al, 2001). Finally, the fossil evidence indicates that the split between placental and marsupial mammals took place more than 125 mya (Colbert et al., 2001). It is important to note that these estimates of time since divergence depend on the preservation of fossil evidence, and given the different habitat types of the four clades it is likely that the quality of the fossil record will differ between groups. Further, fossil evidence in itself only gives the lower limit on lineage age, while the upper limit remains uncertain.

Minimum evolution phylograms based on amino acid (Fig 2.3A) and nucleotide (Fig 2.3B) uncorrected distances show the different levels of divergence between sequences in each of the four clades. The magnitude of pairwise sequence differences observed in the phylograms is predominantly a function of the rate of substitution and time since divergence.

However, the observed divergence may be an underestimate of total divergence if multiple substitutions have affected the same site.

In accordance with the paleontological evidence, the levels of RAG1 sequence divergence among bird taxa are lower than those observed within the other clades and that relationship is unchanged through different partitions and representations of the data (Table 2.2; Fig 2.3). The euteleost sequences show the highest levels of divergence followed by the mammal sequences. There is a significant difference between the highest levels of divergence observed among euteleost and carcharhiniform RAG1 sequences. The maximum divergence among sequences from euteleosts is approximately two times as high as that seen among carcharhiniforms. This difference is observed in all partitions of the data (Table 2.2). Assuming the estimates of age of these clades are approximately correct, then this large sequence divergence difference reflects different rates of evolution and is in agreement with previous reports (Martin et al., 1992).

There are marked differences in divergence levels between the core region and the N-terminal region (Table 2.2). This difference has been noted previously (Gellert, 2001) and probably reflects differences in functional constraints between the two regions. The relationship between divergence measured at nucleotide sites and that measured at amino acid residues differs between the core and N-terminal regions. While nucleotide divergence is higher than amino acid divergence in the core region, the opposite is true in N-terminal regions (Table 2.2). This pattern holds true in all four clades. The changing relationship between nucleotide and amino acid divergence between the core and N-terminal regions indicates that non-synonymous substitutions are less tolerated in the core region causing amino acid divergence to grow at a slower rate than nucleotide divergence. Conversely, in the more variable N-terminal region non-synonymous substitutions contribute

disproportionately to amino acid divergence values (e.g. in a sequence of 100 amino acids, 10 non-synonymous substitutions may result in a 10% amino acid divergence and only a 3.3% nucleotide divergence). Given the high degree of amino acid conservation in the core region and the results of mutation studies in this region, it is likely that negative selection is the prevailing force influencing the fate of sequence variation. The evolutionary dynamics of the N-terminal region appear more complex and heterogeneous. In this region, blocks with high amino acid conservation are interspersed among regions with no recognizable sequence homology between clades (see below). While negative selection probably shapes the conserved blocks, the variable regions could be under neutral or positive selection. The short length of these blocks and the large divergence between the species examined in this study prevent tests to discern between these two possibilities.

All likelihood ratio tests to determine whether or not enforcing equal rates of evolution across all lineages in a given clade had a significant detrimental impact on the likelihood fit of the data to the phylogenetic reconstruction rejected the molecular clock hypothesis. This observation had been reported for the passeriform taxa (Barker et al., 2000) and it is extended here to include three other vertebrate clades. Tests in which the entire sequence was considered and tests restricted to core region sequences produced the same results. Given the phylogenetic relationships between the four clades, it is clear that the euteleostean sequences are evolving at a faster rate than those from the other clades (Fig 3). There is no phylogenetic pattern to the observed differences in rates of change within clades.

*Sequence composition and codon usage bias*

The proportion of nucleotides in a gene sequence is, of course, in part a function of the amino acid composition of the protein product; however, synonymous codons allow the

base composition of genes to vary widely even in the absence of significant changes in amino acid composition. Regardless, if synonymous codons were used essentially at random, the base composition of a gene sequence would not vary significantly between different lineages unless as a result of changes in amino acid composition. The evidence shows that in fact, synonymous codons are rarely used at random and that comparing the base composition at third codon position sites between different lineages often reveals changing biases. Understanding the extent and nature of these biases is important because differences in base composition between lineages are thought to adversely affect the ability of many methods of DNA-based phylogenetic inference to accurately reconstruct evolutionary relationships (Lockhart et al., 1994; but see Rosenberg and Kumar, 2003).

There are significant differences in base composition between and within the groups of RAG1 sequences examined in this study. Each group exhibits distinct patterns of variation (Table 2.3 and Fig 2.4). The group of least divergent sequences, the passeriforms, shows homogenous base composition even when 3$^{rd}$ codon position sites are considered alone. Similarly, the null hypothesis of base composition homogeneity cannot be rejected for the carcharhiniform sequences as a whole or at 3$^{rd}$ codon position sites; however plotting the proportion of the four nucleotides at these sites for each species suggests greater variation among species of sharks than among passeriforms (compare Fig 2.4A with Fig 2.4C). Base composition homogeneity at 3$^{rd}$ codon positions is rejected for both euteleost and mammalian sequences. The base composition heterogeneity at euteleost sequence 3$^{rd}$ codon position sites is so extreme that it results in significant deviation in stationarity when all sites are considered even though 1$^{st}$ and 2$^{nd}$ codon position sites show no significant differences in base composition. All of the esociform sequences are characterized by a similar base composition. The most different patterns of base composition among the euteleost sequences

sampled are represented by *Oncorhynchus mykiss* and *Brachydanio rerio* (Fig 2.4B). Among mammalian sequences, those from *Sus scrofa* and *Lamna glama* show a pattern of base composition that is distinct from all others (Fig 2.4D). It is interesting to note, that the base composition of the sequence from the marsupial *Monodelphis domestica* is more similar to that of placental mammal species than those from the placental *S. scrofa* and *L. glama*. Thus, indicating that differences in base composition bias are not merely the result of incremental divergence with age, but rather must be produced by changes in the dynamics of nucleotide substitution.

Five of the six possible between-group $\chi^2$ tests of base composition reject the null hypothesis of stationarity when all sites are considered (Table 2.3). Remarkably, the exception is the comparison between carcharhiniform and passeriform sequences. When similar comparisons are performed, but only considering $2^{nd}$ codon position sites, the null hypothesis of homogeneity cannot be rejected in five of the six comparisons; while it is rejected for the comparison between euteleost and passeriform sequences (Table 2.3). These last comparisons show that shifts in amino acid composition are not a factor in most of the base composition differences observed. Therefore, the base composition differences between lineages must be the result of evolutionary processes other than selection acting on the biochemistry of the protein product. However, the fact that the sequence composition differences observed above are of greater magnitude when only silent sites are used in the calculation (data not shown) suggests that functional constrains on the gene product limit the impact of substitutional bias on sequence composition.

Lineage dependent biases in codon usage and base composition can be attributed to at least two different forces. First and most relevant to the present study, if the 12 types of nucleotide changing mutations occur or are fixed at unequal frequencies, then this mutational

or substitutional bias can affect the composition of the gene sequence (Singer and Hickey, 2000). But, if mutational bias is to explain base composition and codon usage bias, then its mechanism cannot be based on the biochemical properties of the nucleotides as these are constant across lineages while base composition and codon usage are not. A second force that can affect base composition and codon usage is natural selection, which may act on two different aspects of sequence composition variability. Because base composition can affect the thermal stability of the DNA double helix, it has been suggested that natural selection may act on that variability in stability if the physiology or environment of a species require it (Galtier and Lobry, 1997; but see Wang and Hickey, 2002). Another potential effect of selection on base composition results from expected differences in transcription efficiency between alleles that preferentially contain codons recognized by the most abundant tRNA species and those that exhibit no preferences between synonymous codons. Among vertebrates, thermal stability and efficiency of transcription are unlikely to be important determinants of base composition. Mutational (or substitutional) bias and amino acid composition changes are more likely causes of shifts in gene base composition between vertebrate lineages.

The presence of introns in the euteleost RAG1 sequences provides an opportunity to compare the nucleotide composition of coding and non-coding sequences at that genomic location. Visual inspection and a $\chi^2$ test show that the base composition characteristics of intron and exon sequences at the RAG1 locus are very different ($p<<0.001$; Fig 2.4E). In general, the intron sequences have a more uneven proportion of nucleotides than exon sequences. Further, there is an inverse relationship in the proportions of each nucleotide in intron and exon sequences. For example, in exon sequences guanines are the most abundant nucleotides; while they are least abundant in intron sequences. This difference also applies

to comparisons between introns and silent sites on the exons. These observations are puzzling because they suggest differences in the substitution bias of intron and exon sequences. If it is assumed that changes at silent sites of coding sequences are neutral or nearly so, then the dramatic composition differences between introns and silent sites must reflect functional constraints on the introns. Conversely, if it is assumed that intron sequences have limited evolutionary constraints and therefore their nucleotide composition reflects a lineage-specific and/or locus-specific equilibrium base composition, then the observed differences suggest that substitutions at silent sites are subject to some form of selection.

Another approach to understanding the relationship between substitutional bias and base composition is to examine the deviation from randomness in synonymous codon use. If substitutional bias is the primary factor shaping base composition then codon preferences between different groups of synonymous codons must be similar. For example, if substitutional bias favors increasing proportions of adenines, then the codons with an adenine at its third position should be present in the sequence at a disproportionately high frequency when compared to its synonymous triplets. Comparing the usage of four-fold synonymous codons in the RAG1 sequences in this study reveals some interesting similarities and differences between the four clades examined (Fig 2.5).

The most striking pattern is that codon preference is not the same between different codon groups. For example, in carcharhiniforms there appears to be a bias against the guanine or cytosine ending codons among Proline (CCN) and Alanine (GCN) coding triplets, but the same bias is not evident in the frequency of use of Glycine (GGN) and Valine (GTN) codons (Fig 2.5A). Similar contrasts can be observed in the codon usage in the sequences from the other three clades (Fig 2.5B-D). The clearest contrast occurs in the passeriform

sequences, where the guanine ending codons for Proline, Alanine and Thrionine (ACN) are rare or absent, but well represented among Glycine and Valine coding triplets (Fig 2.5C). This observation suggests that substitution bias is not the sole factor driving codon usage bias. Within each clade, codon usage varies to a moderate extent and the codon preferences of the two taxa with the most dissimilar base composition remain comparable (Fig 2.5).

The effective number of codons (Nc) is another informative measure of bias in codon use. (Wright, 1990) Under random use of synonymous codons, all codons are used at approximately equal frequency and the Nc value approaches the number of residue coding codons in the appropriate genetic code (e.g. 61 for the universal genetic code). When codon usage biases result in underrepresented codons, the Nc has a lower value. The extent of the bias is inversely correlated with the value of Nc. Plotting the Nc index versus the proportion of guanines and cytosines at third codon position silent sites (GC3s) illustrates the differences in base composition and codon usage bias described above (Fig 2.6; Table 2.4). Most mammalian and euteleost sequences favor Gs and Cs, while the carcharhiniform and passeriform sequences have the opposite bias. The effective number of codons for both pairs of clades are similar suggesting a similarly strong codon usage bias although different in the nucleotides that are favored or disfavored.

Between clades, codon usage is most similar between mammalian and euteleost sequences. The codon usage of carcharhiniform and passeriform sequences are similar to a lesser degree. The codon usage differences and similarities between clades correspond well with those seen for base composition suggesting a common causes underlying these shifting sequence characteristics. And, given that substitution bias alone does not appear to be responsible for codon usage bias, then base composition itself may not be the consequence of substitutional bias.

*Conservation and change along the RAG1 protein*

In agreement with the complex picture of the functional anatomy of RAG1 that is emerging from functional studies, the comparative approach utilized here shows that the RAG1 protein consists of several of blocks with high amino acid residue conservation connected by highly variable regions that tolerate both amino acid substitution and length altering mutations. Following is a description of the conservation characteristics of the RAG1 sequences examined in the context of the known roles of the different parts of the molecule.

At the N-terminal end of RAG1, there is a short highly conserved motif that extends between residues 15 and 40 of the *Mus musculus* sequence (Fig 2.7). The motif is restricted to the bony vertebrate clades and no region with recognizable homology to this motif is present in the carcharhiniform sequences, which could indicate that the motif is a novel feature of the bony vertebrate lineage or that it has been lost from the carcharhiniform protein. At the core of the motif is the sequence WKFKLF(R/K)V, which is unchanged in all known bony vertebrate sequences that cover this region (n=14). Clearly there is a functional constraint operating to preserve the precise identity of this motif for over 400 million years; however no role has been ascribed to it to date. Searches of publicly available databases of functional motifs and nuclear localization signals failed to produce significant matches. The motif is richer in aromatic and basic residues than the regions that flank it.

Following the motif described above, there is a variable length region spanning anywhere from 64 to 105 residues located between position 41 and 108 of the *M. musculus* sequence. For this region, alignment is possible and homology recognizable only within major vertebrate clades. Between clades the alignment is problematic and there are no conserved residues that span multiple clades. It is unclear whether the contrasting patterns of

conservation in this region within and between clades is the product of the shorter time since divergence of the sequences within clades or of divergent functional constraints operating between clades.

The variable region is then interrupted by another conserved sequence block. This conserved region extends between residues 109 and 213 of the *M. musculus* sequence (Fig 2.8). Seventeen residues interspersed in this block are conserved in all the sequences examined and many others are conserved in the majority but not all of the sequences. Six sites exhibit conservative amino acid substitutions. Among the completely conserved residues there are three widely spaced cysteine (Cys) pairs. In all three pairs and all the sequences available, two residues separate the Cys residues: CXXC, which takes the form CRXC, FCXXCW and CX(ILV)C in the first, second and third pairs, respectively. The conserved two-residue spacing between the conserved Cys residues in this block resembles the ion coordinating Cys of zinc fingers, although the three pairs span a greater length (over 100 residues) than the two zinc fingers (ZFA and ZFB) identified in RAG1 to date and, with the exception of a conserved histidine (His) found 15 positions after the first Cys pair in the tetrapod sequences, there are no appropriately placed Histidines following the Cys motif. There is a conserved tryptophan (Trp) located between the first and second Cys pairs and another one between the second and third pairs. The spacing between the first two Cys pairs is highly conserved with length variation limited to a one-residue deletion in all of the euteleostean sequences. The spacing between the second and third Cys pairs is less conserved. There are five-residue and one-residue deletions in the euteleostean and tetrapod sequences, respectively. Amino acid identity and property conservation in the sequences between Cys pairs is low between clades. Searches of protein motif databases produce weak and partial matches to zinc finger motifs (e.g. GATA zinc finger). The conserved spacing of

conserved residues in this region and the complete conservation of the Cys pairs strongly suggest a functional constraint, however no role has been ascribed to these residues. The first of five basic motifs (BI) is located between the first and second Cys pairs. In tetrapods, BI consists of four or five basic residues in a stretch spanning six residues. In carcharhiniforms and euteleosts, the BI homologous region contains only two and three basic residues, respectively. A role in the nuclear localization of RAG1 has been suggested for BI (Cortes et al., 1994). As nuclear localization is essential for V(D)J recombination then there must be an alternative basic motif in the carcharhiniform sequence, or the role of BI is not essential for nuclear localization.

Following the block containing the CXXC pairs, there is a variable region between positions 214 and 265 of the *M. musculus* sequence with significant length differences between euteleostean and all other sequences. This region spans up to 75 residues in euteleosts and close to 50 in all other sequences examined. Interestingly, there are two completely conserved Trp residues and two highly conserved Trp residues in this region of the euteleostean sequences. None of these are present in the other clades. This region is rich in basic amino acids and it contains basic motifs BII and BIII. The residues homologous to BII contain a short stretch of basic amino acids in all of the sequences examined. By contrast, BIII is not immediately recognizable in euteleostean or passeriform sequences. In these clades, there are a number of basic residues interspersed in the region, but not the contiguous stretch of such amino acids that characterizes the basic motif.

The dimerization domain follows the length variable region. The dimerization domain extends between positions 266 and 380 and contains the RING and Zinc finger domains (Bellon et al., 1997; Fig 2.9). There is a high degree of conservation and no length variation in this region. Twenty-one sites, including the 15 ion coordinating Cys and His

residues of the finger domains, are conserved in all sequences examined. Two of the three Phe residues (284 and 344) that are thought to form the hydrophobic cores of the dimerization interface are highly conserved. Phe 280 is replaced by tyrosine in carcharhiniforms and non-therian tetrapods, Trp in *Lama glama* and leucine and valine in euteleosts. All of these substitutions preserve the hydrophobic quality of the dimerization interface.

In carcharhiniforms and tetrapods, the core region closely follows the dimerization domain with only nine sites separating the last and the first highly conserved residues in each domain (positions 376 and 386 of the *Mus musculus* sequence in the dimerization and core regions, respectively). In euteleost sequences, 19 to 21 residues separate these two domains. Further, the short intervening sequence in carcharhiniforms and tetrapods is rich in basic amino acids, while its longer homologue in euteleosts is not.

As expected, there is only limited length variation in the core domain and sequence conservation is remarkable (Fig 2.10). Only a few insertions and deletions are found within the part of the core region that has been implicated in nonamer binding (positions 389 to 445 of *M. musculus*). Among the sequences examined, 280 of the 587 core region sites for which all sequences are represented, are fully conserved. These include the five catalytic sites and 26 of the 27 basic amino acid mutants that are known to be severely defective in recombination assays (Huye et al., 2002). The exception occurs at position 839 where a lysine is replaced by glutamine and threonine in the esociform fish taxa and by glutamine and glutamic acid in three different carcharhiniforms. This site falls within BIV, which is less conserved in euteleosts than in the other clades. All four basic amino acids of BV are conserved across all sequences examined, which is consistent with its dominant role in nuclear localization of RAG1 (Spanopoulou et al., 1995).

To summarize, an examination of patterns of change and conservation along the RAG1 gene revealed that there are several highly conserved features for which functions remain to be determined. These are: the short motif located in the N-terminus of the protein in all bony vertebrate sequences known, but with no identifiable homologue in sharks; the block that contains three fully conserved Cys pairs and intervening Trp; and the lengthened region between the Cys pair block and dimerization domain distinguishing the euteleost sequences from all other known sequences. A schematic representation of the conserved sequence blocks and intervening variable regions illustrate some of the structural complexity of the RAG1 protein that reflects its functional complexity (Fig 2.11).

*Implications for phylogenetic analysis based on RAG1 sequences*

One of the most problematic issues in phylogenetic reconstruction is the need to accommodate the heterogeneous evolutionary forces that shape molecules. Under many methods of phylogenetic inference, the investigator can try to capture this heterogeneity by adding parameters that estimate different characteristics of sequence evolution. As methods increase in sophistication the number of parameters that can be considered grows but outside of studies based on simulated data, it is not always clear how to chose an appropriate parameterization given a set of sequences. One reasonable approach is to examine the sequence variation characteristics of different genes in search of sequences that evolve in a manner that fit the simpler models employed by reconstruction methods. One such analysis conducted by G. Naylor (personal communication) suggested that the variability characteristics of RAG1 sequences held potential for phylogenetic reconstruction. The present study further examines the suitability of RAG1 sequences for phylogenetic inference using a more diverse sample of sequences.

In agreement with G. Naylor's analysis, the RAG1 sequences examined in this study show some promising qualities for phylogenetic inference. First, there is no evidence of duplication in any of the taxa examined to date. Second, the constraints placed on the molecule by its important and complex function limit many aspects of its evolution. This is desirable in phylogenetic inference because the effects of functional constraints make multiple sequence alignment relatively easy over most of the length of the molecule, thus giving confidence on the homology relationships between aligned sites in different species. Another consequence of functional constraint is a relatively slow rate of change, which gives this gene the potential to preserve phylogenetic information over a longer time than the commonly used mitochondrial genes.

On the other hand, there are several characteristics of this gene that may violate some of the assumptions implicit in many methods of phylogenetic reconstruction. Perhaps the most important of these is the varied patterns of evolution across different parts of the molecule. Its highly conserved domains and intervening variable regions show very different rates and types of substitution. For example, insertions and deletions are rare in conserved regions and common in variable regions. This heterogeneous evolutionary dynamics make the alignment of variable regions problematic. The alignment problems may only be relevant to those studies whose aim is to infer relationships between vertebrate classes. Fortunately, the proportion of highly variable regions is small, so it may be desirable to disregard them, or at least treat with caution, in the investigation of the most ancient vertebrate divergences. In studies that focus on more recent evolutionary relationships (e.g. within vertebrate orders), alignment of variable regions may not be problematic. However, they still represent a subset of the data with unique evolutionary constraints, which need to be accommodated by the methods of inference.

Even when variable regions are disregarded there is heterogeneity in the degree of conservation between conserved sequence blocks. The 'CXXC' and dimerization domains exhibit less sequence conservation than the core region. Within those domains, residue identity conservation is mostly restricted to a small proportion of residues, while almost half of all the sites in the core region are fully conserved. These differences in the evolutionary dynamics of the conserved regions indicate that there will be variation in the absolute and relative rate of synonymous and non-synonymous substitutions among regions of the gene sequence. Therefore, the method of phylogenetic inference must consider this type of variation or, if not, at least be robust to its presence.

Base composition and codon usage in the RAG1 gene show statistically significant differences between different clades and, to a more limited extent, within the euteleost and mammalian clades. In most cases, base composition heterogeneity is limited to third codon position sites. Base composition heterogeneity has been shown to adversely affect the accuracy of phylogenetic reconstruction (Lockhart et al., 1994), although recent studies based on simulated sequence data suggest that the effect may be negligible (Rosenberg and Kumar, 2003). Nevertheless, when drawing phylogenetic inferences from RAG1 sequences, it is advisable to examine the base composition characteristics of the groups of interest. This is especially true when the conclusions of the phylogenetic analysis support suspect groupings.

The RAG1 sequences examined here do not follow a clock-like rate of evolution between or within clades. While most methods of phylogenetic inference can accommodate some rate variation, the combined effects of variable rates and poor taxonomic sampling can lead to erroneous inference. Therefore, the among-lineage rate variability of RAG1 sequences requires that long branches in a taxonomic sample be avoided to the extent that the extant diversity of the group of interest permits it. Unfortunately, taxonomic sampling has an

upper limit dictated by the extinction history of a given group, thus inferences of relationships of poorly sampled or species-poor clades based on RAG1 sequences may be adversely affected by long-branch attraction (Felsenstein, 1978). Another negative consequence of the prevalent variation in rate of evolution between lineages is that this gene may not be useful for the estimation of the age of different lineage formation events.

In conclusion, RAG1 sequences possess many characteristics that represent an improvement over other genes commonly used in phylogenetic reconstruction (notably, mitochondrial DNA and nuclear ribosomal RNA). But at the same time, the functional complexity of the RAG1 protein and its shifting substitution bias and rate of change require that the investigator explore the interplay between this variability and the inferences supported by different methods. Unfortunately, the evolution of the RAG1 gene does not seem to fit the simpler models implemented by current methods of phylogenetic inference, therefore its use in such studies requires the same combination of experience and guesswork that is characteristic in the field.

Finally, the intron-exon structure of the RAG1 gene has an interesting phylogenetic distribution. An extensive sample of RAG1 sequences from sharks reveals an intronless gene without exception (G. Naylor, personal communication). Similarly, all tetrapod sequences available to date, show the absence of introns in the RAG1 gene of this lineage (e.g. Barker et al., 2002). These observations support the view that the ancestral RAG1 gene in vertebrates was devoid of introns and consequently that the introns found in the RAG1 gene in actinopterygian fishes originated after the divergence between the actinopterygian and sarcopterygian lineages. Further sampling of the distribution of the four different RAG1 introns among actinopterygian lineages may prove to be phylogenetically informative (Venkatesh et al., 1999; López et al., accepted). Using the presence or absence of introns

and conserved insertions and deletions (Venkatesh et al., 2001) as sources of phylogenetic information obviates many of the problems associated with the need to capture the complex evolutionary dynamics of gene sequences. However, the relative rarity of these character types when compared to nucleotide or amino acid sites negates one of the advantages molecular techniques brings to phylogenetic systematics: a practically limitless amount of relevant data.

**Acknowledgements**

I thank the member of the Naylor Laboratory for providing the shark RAG1 sequences, as well as the information on RAG1 amplification and sequencing used to obtain some of the sequences from fishes.

**References**

Agrawal A, Schatz DG (1997) Rag1 and Rag2 Form a Stable Postcleavage Synaptic Complex With Dna Containing Signal Ends in V(D)J Recombination. Cell 89:43-53

Aidinis V, Dias DC, Gomez CA, Bhattacharyya D, Spanopoulou E, Santagata S (2000) Definition of minimal domains of interaction within the recombination-activating genes 1 and 2 recombinase complex. J Immunol 164:5826-5832

Arbuckle JL, Fauss LA, Simpson R, Ptaszek LM, Rodgers KK (2001) Identification of two topologically independent domains in RAG1 and their role in macromolecular interactions relevant to V(D)J recombination. J Biol Chem 276:37093-37101

Banerjee-Basu S, Baxevanis AD (2002) The DNA-binding region of RAG 1 is not a homeodomain. Genome Biol 3:interactions1004.1-1004.4

Barker FK, Barrowclough GF, Groth JG (2002) A phylogenetic hypothesis for passerine birds: Taxonomic and biogeographic implications of an analysis of nuclear DNA sequence data. Proc Royal Soc Biol Sci Series B 269: 295-308

Barton GJ (1993) ALSCRIPT - A tool to format multiple sequence alignments. Protein Eng 6:37-40

Bellon SF, Rodgers KK, Schatz DG, Coleman JE, Steitz TA (1997) Crystal Structure of the Rag1 Dimerization Domain Reveals Multiple Zinc-Binding Motifs Including a Novel Zinc Binuclear Cluster. Nat Struct Biol 4:586-591

Bernstein RM, Schluter SF, Bernstein H, Marchalonis JJ (1996) Primordial emergence of the recombination activating gene 1 (RAG1): sequence of the complete shark gene indicates homology to microbial integrases. Proc Natl Acad Sci USA 93:9454-9459

Ciubotaru M, Ptaszek LM, Baker GA, Baker SN, Bright FV, Schatz DG (2003) RAG1-DNA binding in V(D)J recombination. Specificity and DNA-induced conformational changes revealed by fluorescence and CD spectroscopy. J Biol Chem 278:5584-5596

Cortes P, Ye ZS, Baltimore D (1994) Rag-1 Interacts With the Repeated Amino Acid Motif of the Human Homologue of the Yeast Protein Srp1. Proc Natl Acad Sci USA 91:7633-7637

Cuomo CA, Kirch SA, Gyuris J, Brent R, Oettinger MA (1994) Rch1, a Protein That Specifically Interacts With the Rag-1 Recombination-Activating Protein. Proc Natl Acad Sci USA 91:6156-6160

Dai Y, Kysela B, Hanakahi LA, Manolis K, Riballo E, Stumm M, Harville TO, West SC, Oettinger MA, Jeggo PA (2003) Nonhomologous end joining and V(D)J recombination require an additional factor. Proc Natl Acad Sci USA 100:2462-2467

Davis MM (1990) T cell receptor gene diversity and selection. Annu Rev Biochem 59:475-496

Feduccia A (1996) The origin and evolution of birds. Yale University Press, New Haven

Felsenstein J (1978) Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool 27:401-410

Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperatures in prokaryotes. J Mol Evol 44:632-636

Gellert M (2002) V(D)J recombination: RAG proteins, repair factors, and regulation. Annu Rev Biochem 71:101-132

Godderz LJ, Rahman NS, Risinger GM, Arbuckle JL, Rodgers KK (2003) Self-association and conformational properties of RAG1: implications for formation of the V(D)J recombinase. Nucleic Acids Res 31:2014-2023

Hedges SB, Parker PH, Sibley CG, Kumar S (1996) Continental breakup and the ordinal diversification of birds and mammals. Nature 381:226-229

Huye LE, Purugganan MM, Jiang MM, Roth DB (2002) Mutational analysis of all conserved basic amino acids in RAG-1 reveals catalytic, step arrest, and joining-deficient mutants in the V(D)J recombinase. Mol Cell Biol 22:3460-3473

Landree MA, Wibbenmeyer JA, Roth DB (1999) Mutational analysis of RAG1 and RAG2 identifies three catalytic amino acids in RAG1 critical for both cleavage steps of V(D)J recombination. Genes Dev 13:3059-3069

Livak F, Petrie HT (2002) Access roads for RAG-ged terrains: control of T cell receptor gene rearrangement at multiple levels. Semin Immunol 14:297-309

Livingstone CD, Barton GJ (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. CABIOS 9:745-756

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605-612

Lopez JA, Bentzen P, T.W. P (2000) Phylogenetic relationships of esocoid fishes (Teleostei) based on partial cytochrome b and 16S mitochondrial DNA sequences. Copeia 2000:420-431

Martin AP, Naylor GJ, Palumbi SR (1992) Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. Nature 357:153-155

McMahan CJ, Difilippantonio MJ, Rao N, Spanopoulou E, Schatz DG (1997) A Basic Motif in the N-Terminal Region of Rag1 Enhances V(D)J Recombination Activity. Mol Cell Biol 17:4544-4552

Minkoff EC (2001) Colbert's evolution of the vertebrates: a history of the backboned animals through time. John Wiley, New York

Mo X, Bailin T, Sadofsky MJ (2001) A C-terminal region of RAG1 contacts the coding DNA during V(D)J recombination. Mol Cell Biol 21:2038-2047

Moshous D, Callebaut I, de Chasseval R, Corneo B, Cavazzana-Calvo M, Le Deist F, Tezcan I, Sanal O, Bertrand Y, Philippe N, Fischer A, de Villartay JP (2001) Artemis, a novel DNA double-strand break repair/V(D)J recombination protein, is mutated in human severe combined immune deficiency. Cell 105:177-186

Nakase H, Takahama Y, Akamatsu Y (2003) Effect of CpG methylation on RAG1/RAG2 reactivity: implications of direct and indirect mechanisms for controlling V(D)J cleavage. EMBO Rep. 4:774-780

Noordzij JG, Verkaik NS, Hartwig NG, de Groot R, van Gent DC, van Dongen JJ (2000) N-terminal truncated human RAG1 proteins can direct T-cell receptor but not immunoglobulin gene rearrangements. Blood 96:203-209

Peak MM, Arbuckle JL, Rodgers KK (2003) The central domain of core RAG1 preferentially recognizes single-stranded recombination signal sequence heptamer. J Biol Chem 278:18235-18240

Rodgers KK, Bu ZM, Fleming KG, Schatz DG, Engelman DM, Coleman JE (1996) A Zinc-Binding Domain Involved in the Dimerization of Rag1. J Mol Biol 260:70-84

Rodgers KK, Villey IJ, Ptaszek L, Corbett E, Schatz DG, Coleman JE (1999) A dimer of the lymphoid protein RAG1 recognizes the recombination signal sequence and the complex stably incorporates the high mobility group protein HMG2. Nucleic Acids Res 27:2938-2946

Rosenberg MS, Kumar S (2003) Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol Biol Evol 20:610-621

Roth DB, Roth SY (2000) Unequal access: Regulating V(D)J recombination through chromatin remodeling. Cell 103:699-702

Sadofsky MJ (2001) The RAG proteins in V(D)J recombination: more than just a nuclease. Nucleic Acids Res 29:1399-409

Sadofsky MJ, Hesse JE, McBlane JF, Gellert M (1993) Expression and V(D)J Recombination Activity of Mutated Rag-1 Proteins. Nucleic Acids Res 21:5644-5650

Sekiguchi JA, Whitlow S, Alt FW (2001) Increased accumulation of hybrid V(D)J joins in cells expressing truncated versus full-length RAGs. Mol Cell 8:1383-1390

Shih IH, Melek M, Jayaratne ND, Gellert M (2002) Inverse transposition by the RAG1 and RAG2 proteins: role reversal of donor and target DNA. EMBO J 21:6625-6633

Shirai S (1996) Phylogenetic relationships of Neoselachians. In: Stiassny MLJ, Parenti L, Johnson GD (eds) Interrelationships of fishes. Academic Press, San Diego, pp 9-32

Singer GA, Hickey DA (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol 17:1581-1588

Spanopoulou E, Cortes P, Shih C, Huang CM, Silver DP, Svec P, Baltimore D (1995) Localization, Interaction, and Rna Binding Properties of the V(D)J Recombination-Activating Proteins Rag1 and Rag2. Immunity 3:715-726

Spanopoulou E, Zaitseva F, Wang FH, Santagata S, Baltimore D, Panayotou G (1996) The Homeodomain Region of Rag-1 Reveals the Parallel Mechanisms of Bacterial and V(D)J Recombination. Cell 87:263-276

Steen SB, Han JO, Mundy C, Oettinger MA, Roth DB (1999) Roles of the "dispensable" portions of RAG-1 and RAG-2 in V(D)J recombination. Mol Cell Biol 19:3010-3017

Swanson PC (2002) A RAG-1/RAG-2 tetramer supports 12/23-regulated synapsis, cleavage, and transposition of V(D)J recombination signals. Mol Cell Biol 22:7790-7801

Swofford DL (1998) PAUP*. Phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Sunderland, Massachussetts

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876-4882

van Gent DC, Mizuuchi K, Gellert M (1996) Similarities between initiation of V(D)J recombination and retroviral integration. Science 271:1592-1594

Venkatesh B, Ning Y, Brenner S (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. Proc Natl Acad Sci USA 96:10267-10271

Waddell PJ, Shelley S (2003) Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus

MCMC-driven nucleotide, amino acid, and codon models. Mol Phylogenet Evol 28:197-224

Wang HC, Hickey DA (2002) Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes. Nucleic Acids Res 30:2501-2507

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23-29

Yurchenko V, Xue Z, Sadofsky M (2003) The RAG1 N-terminal domain is an E3 ubiquitin ligase. Genes Dev 17:581-585

**Table 2.1.** RAG1 functions related to V(D)J recombination

| Function | Residues Implicated | References |
|---|---|---|
| 1. Interaction with other factors: | | |
| a. Subcellular localization (nuclear, intranuclear): | | |
| i. Interaction with importin α 1 subunit | 141-146, 221-224, 243-249 | Cortes et al., 1994 |
| ii. Interaction with importin α 2 subunit | 826-840, 969-973 | Cuomo et al., 1994 |
| b. RAG1Dimerization: | | |
| i. Ring and Zinc (ZFA) fingers | 266-330, 352-378 | Rodgers et al., 1996<br>Bellon et al., 1997 |
| ii. C-terminal domain | 761-979 | Arbuckle et al., 2001 |
| c. Binding to RAG2: | | |
| i. Zinc finger B (ZFB) | 726-750 | Aidinis et al., 2000 |
| d. Undetermined interaction with V(D)J factors | | |
| e. Protein degradation: | | |
| i. E3 ubiquitin ligase (RING finger) | 266-330 | Yurchenko et al., 2002 |

**Table 2.1.** (continued)

| Function | Residues Implicated | References |
|---|---|---|
| 2. V(D)J reaction: | | |
| a. RSS recognition and binding: | | |
| i. Heptamer | 528-760 | Peak et al., 2003 |
| ii. Nonamer | 384-477 | Spanopoulou et al., 1996 |
| iii. Coding sequence | 889-974 | Mo et al., 2001 |
| b. Catalytic residues: | | |
| i. Hydrolysis | 600, 708, 713, 795, 962 | Landree et al., 1999 |
| | | Huye et al., 2002 |

**Table 2.2.** Range of percent pairwise sequence divergence values observed among members of each of the four clades examined in this study. The values given are the uncorrected percentage of sites differing between a pair of sequences.

| Group | Overall sequence | | | Core region[1] | | N-terminal region[2] | |
|---|---|---|---|---|---|---|---|
| | AA[3] | All nucs.[4] | 3rd pos.[5] | AA | All nucs. | AA | All nucs. |
| Carcharhiniformes | 1.4-11.9 | 1.5-14.7 | 2.3-31.2 | 0.8-9.8 | 1.7-14.1 | 1.6-21.9 | 2.3-17.4 |
| Euteleostei | 0.8-24.4 | 1.1-28.8 | 2.4-58.6 | 0.5-17.0 | 0.8-25.5 | 1.5-45.9 | 1.3-40.7 |
| Passeriformes[6] | 1.4-7.6 | 2.5-10.3 | 6.1-22.9 | 0.5-3.9 | 1.9-9.2 | 3.2-18.5 | 2.8-15.0 |
| Mammalia | 2.5-15.6 | 6.3-22.1 | 18.3-46.3 | 1.6-8.0 | 6.0-18.6 | 6.5-38.5 | 7.2-31.9 |

[1] Sites corresponding to amino acid residues 384-1008 of the *Mus musculus* protein.

[2] Sites corresponding to amino acid residues 1-383 of the *M. musculus* protein. Due to length variation in this region of the molecule the number of sites considered is different for each group.

[3] Divergence of the translated sequences.

[4] Divergence at all nucleotide sites within the corresponding region.

[5] Divergence at 3rd codon position sites within the corresponding region.

[6] Includes comparisons of passerine species with the outgroup *Gallus gallus*.

**Table 2.3.** P-values of $\chi^2$ tests of base composition homogeneity of the RAG1 gene sequences

| | Within groups[1] | | Between groups[2] | | | |
|---|---|---|---|---|---|---|
| | All nucs.[3] | 3rd pos.[4] | Carcharhiniformes | Euteleostei | Passeriformes | Mammalia |
| Carcharhiniformes | 0.997[5] | 0.460 | | < 0.001**[6] | 0.999 | < 0.001** |
| Euteleostei | < 0.001** | < 0.001** | 0.312[7] | | < 0.001** | < 0.001** |
| Passeriformes | 0.999 | 0.999 | 1.0 | 0.005* | | < 0.001** |
| Mammalia | 0.045* | < 0.001** | 1.0 | 0.313 | 1.0 | |

[1] $\chi^2$ test limited to sequences from taxa within each of the four vertebrate clades examined.

[2] $\chi^2$ test including sequences from taxa belonging to two of the four vertebrate clades examined.

[3] Nucleotides at all sites considered in the $\chi^2$ test.

[4] Only nucleotides at 3rd codon position sites considered in the $\chi^2$ test.

[5] P-value of the $\chi^2$ tests. * highlights values below 0.05; ** highlights values below 0.001.

[6] Values above the diagonal are from $\chi^2$ tests overall all nucleotide sites.

[7] Values below the digonal are from $\chi^2$ tests for 2nd codon position sites.

**Table 2.4.** Base composition at silent sites (N3s) and effective number of codons (Nc)

| Species[1] | A3s[2] | C3s | G3s | T3s | GC3s | Nc[3] |
|---|---|---|---|---|---|---|
| Carcharhiniformes | | | | | | |
| Aprlab | 0.222 | 0.393 | 0.405 | 0.269 | 0.371 | 51.44 |
| Eridac | 0.237 | 0.391 | 0.394 | 0.271 | 0.383 | 52.51 |
| Galboa | 0.231 | 0.400 | 0.391 | 0.269 | 0.378 | 52.62 |
| Rhiacu | 0.255 | 0.397 | 0.373 | 0.268 | 0.396 | 52.62 |
| Galgal | 0.250 | 0.390 | 0.366 | 0.282 | 0.405 | 53.01 |
| Mushen | 0.248 | 0.390 | 0.368 | 0.284 | 0.403 | 53.05 |
| Sphzyg | 0.247 | 0.417 | 0.372 | 0.257 | 0.381 | 53.11 |
| Cardus | 0.251 | 0.399 | 0.371 | 0.271 | 0.395 | 53.54 |
| Negacu | 0.249 | 0.405 | 0.371 | 0.269 | 0.392 | 54.03 |
| Chamac | 0.255 | 0.388 | 0.364 | 0.284 | 0.409 | 54.28 |
| Carleu | 0.251 | 0.398 | 0.370 | 0.269 | 0.396 | 54.38 |
| Lepsmi | 0.278 | 0.376 | 0.356 | 0.285 | 0.426 | 54.84 |
| Porpan | 0.290 | 0.379 | 0.329 | 0.293 | 0.443 | 55.47 |
| Hallin | 0.302 | 0.340 | 0.338 | 0.310 | 0.467 | 55.90 |

[1] Within each clade, the list of species is sorted by the Nc index, in ascending order to more clearly show the range of codon usage bias in each group.

[2] Proportion of adenines at third codon position sites, where adenine substitution would be silent.

[3] Wright, 1990.

**Table 2.4.** (continued)

| Species | A3s | C3s | G3s | T3s | GC3s | Nc |
|---|---|---|---|---|---|---|
| Euteleostei | | | | | | |
| Oncmyk | 0.476 | 0.148 | 0.184 | 0.449 | 0.732 | 44.30 |
| Esorei | 0.441 | 0.184 | 0.215 | 0.419 | 0.679 | 49.11 |
| Esoluc | 0.441 | 0.191 | 0.210 | 0.420 | 0.679 | 49.30 |
| Novhub | 0.423 | 0.193 | 0.229 | 0.413 | 0.660 | 49.63 |
| Esomas | 0.429 | 0.193 | 0.228 | 0.408 | 0.661 | 50.19 |
| Esoame | 0.422 | 0.208 | 0.234 | 0.398 | 0.646 | 50.72 |
| Dalpec | 0.427 | 0.190 | 0.231 | 0.411 | 0.662 | 50.95 |
| Umblim | 0.423 | 0.219 | 0.235 | 0.380 | 0.635 | 51.21 |
| Umbkra | 0.434 | 0.223 | 0.232 | 0.370 | 0.635 | 51.65 |
| Takrub | 0.385 | 0.266 | 0.278 | 0.339 | 0.565 | 54.65 |
| Brarer | 0.340 | 0.301 | 0.319 | 0.312 | 0.507 | 56.27 |
| Passeriformes | | | | | | |
| Embsch | 0.282 | 0.355 | 0.375 | 0.295 | 0.433 | 52.34 |
| Smiruf | 0.258 | 0.383 | 0.390 | 0.278 | 0.399 | 52.46 |
| Stuvul | 0.308 | 0.349 | 0.337 | 0.316 | 0.467 | 52.69 |
| Cerfam | 0.284 | 0.362 | 0.366 | 0.297 | 0.435 | 52.76 |
| Forcol | 0.271 | 0.375 | 0.369 | 0.294 | 0.422 | 53.00 |
| Prucol | 0.285 | 0.349 | 0.371 | 0.308 | 0.442 | 53.02 |
| Tyrtyr | 0.267 | 0.363 | 0.379 | 0.301 | 0.424 | 53.29 |
| Hirpyr | 0.279 | 0.374 | 0.362 | 0.291 | 0.427 | 53.66 |
| Aegiou | 0.293 | 0.368 | 0.356 | 0.291 | 0.438 | 53.74 |

**Table 2.4.** (continued)

| Species | A3s | C3s | G3s | T3s | GC3s | Nc |
|---|---|---|---|---|---|---|
| Picgum | 0.285 | 0.362 | 0.365 | 0.299 | 0.436 | 53.74 |
| Oedili | 0.285 | 0.358 | 0.355 | 0.308 | 0.446 | 54.19 |
| Gallga | 0.301 | 0.380 | 0.356 | 0.280 | 0.432 | 54.38 |
| Parstr | 0.283 | 0.361 | 0.364 | 0.300 | 0.436 | 54.57 |
| Corlin | 0.285 | 0.359 | 0.359 | 0.305 | 0.442 | 54.95 |
| Orilar | 0.297 | 0.351 | 0.342 | 0.314 | 0.460 | 55.50 |
| Mammalia | | | | | | |
| Lamgla | 0.444 | 0.185 | 0.243 | 0.425 | 0.663 | 49.66 |
| Susscr | 0.448 | 0.202 | 0.243 | 0.409 | 0.651 | 49.93 |
| Homsap | 0.360 | 0.251 | 0.335 | 0.355 | 0.541 | 51.87 |
| Orycan | 0.358 | 0.251 | 0.330 | 0.362 | 0.545 | 51.87 |
| Ratnor | 0.375 | 0.251 | 0.317 | 0.354 | 0.555 | 53.85 |
| Mondom | 0.376 | 0.260 | 0.313 | 0.346 | 0.550 | 53.85 |
| Musmus | 0.376 | 0.248 | 0.315 | 0.358 | 0.558 | 54.18 |

**Figure Legends**

**Figure 2.1** A. Structure of recombination signal sequences (RSS's); B. Arrangement of the two RSS types (distinguished by filled and open triangles) on the variable (V), diversity (D) and joining (J) segments of an antigen receptor locus; C. Coding end processing steps leading to a modified coding end prior to recombination.

**Figure 2.2** Relative location of the amplification and sequencing primers used to obtain RAG1 DNA sequences from esociform fishes. Primer locations are marked by arrows on a schematic representation of the RAG1 coding sequence of *Oncorhynchus mykiss*. Filled regions represent introns.

**Figure 2.3** Phylograms of the sequences examined based on uncorrected percent pairwise sequence divergences of (A) amino acid and (B) nucleotides. The phylograms were rooted with the carcharhiniform sequences. Species abbreviations are given in the text. The four clades are: (1) Passeriformes plus *Gallus gallus*, (2) Theria, (3) Euteleostei and (4) Carcharhiniformes.

**Figure 2.4** Proportion of each of the four nucleotides in the RAG1 sequences arranged by degree of similarity. Species abbreviations are given in the text. A. Carcharhiniformes; B. Euteleostei; C. Passeriformes plus *Gallus gallus*; D. Theria; E. Representative species from the four groups; F. Intron (I) and exon sequences of euteleost taxa.

**Figure 2.5** Proportion of each of codon from four-fold synonymous codong groups for the two most dissimilar species in each group. Species abbreviations are given in the text. A. Carcharhiniformes; B. Euteleostei; C. Passeriformes plus *Gallus gallus*; D. Theria.

**Figure 2.6** Plot of the effective number of codons (Nc) versus the proportion of G's and C's at third codon position silent sites.

**Figure 2.7** Alignment of the short conserved block found in the N-terminal region of all bony vertebrate sequences. Species abbreviations are given in the text. Black background indicates identity conservation in all sequences compared; grey background indicates conservation within groups; and, white background indicates variability within groups. The histogram below the alignment indicates the degree of amino acid identity conservation.

**Figure 2.8** Alignment of the conserved block spanning the three conserved pairs of Cys residues: 3(CXXC). Species abbreviations are given in the text. Black background indicates identity conservation in all sequences compared; grey background indicates conservation within groups; and, white background indicates variability within groups. The histogram below the alignment indicates the degree of amino acid identity conservation.

**Figure 2.9** Alignment of the dimerization domain, including the RING and zinc finger (ZFA) domains. Species abbreviations are given in the text. Black background indicates identity conservation in all sequences compared; grey background indicates conservation within groups; and, white background indicates variability within groups. The histogram below the alignment indicates the degree of amino acid identity conservation.

**Figure 2.10** Alignment of the region of core RAG1 overlapped by all the sequences examined. Species abbreviations are given in the text. Black background indicates identity conservation in all sequences compared; grey background indicates conservation within groups; and, white background indicates variability within groups. The histogram below the alignment indicates the degree of amino acid identity conservation.

**Figure 2.11** Schematic diagram of the patterns of conservation along the RAG1 protein. Filled blocks are conserved regions, open blocks are variable regions. Labeled blocks are described in the text. Asterisks indicate regions where insertions and deletions are observed between different vertebrate clades.

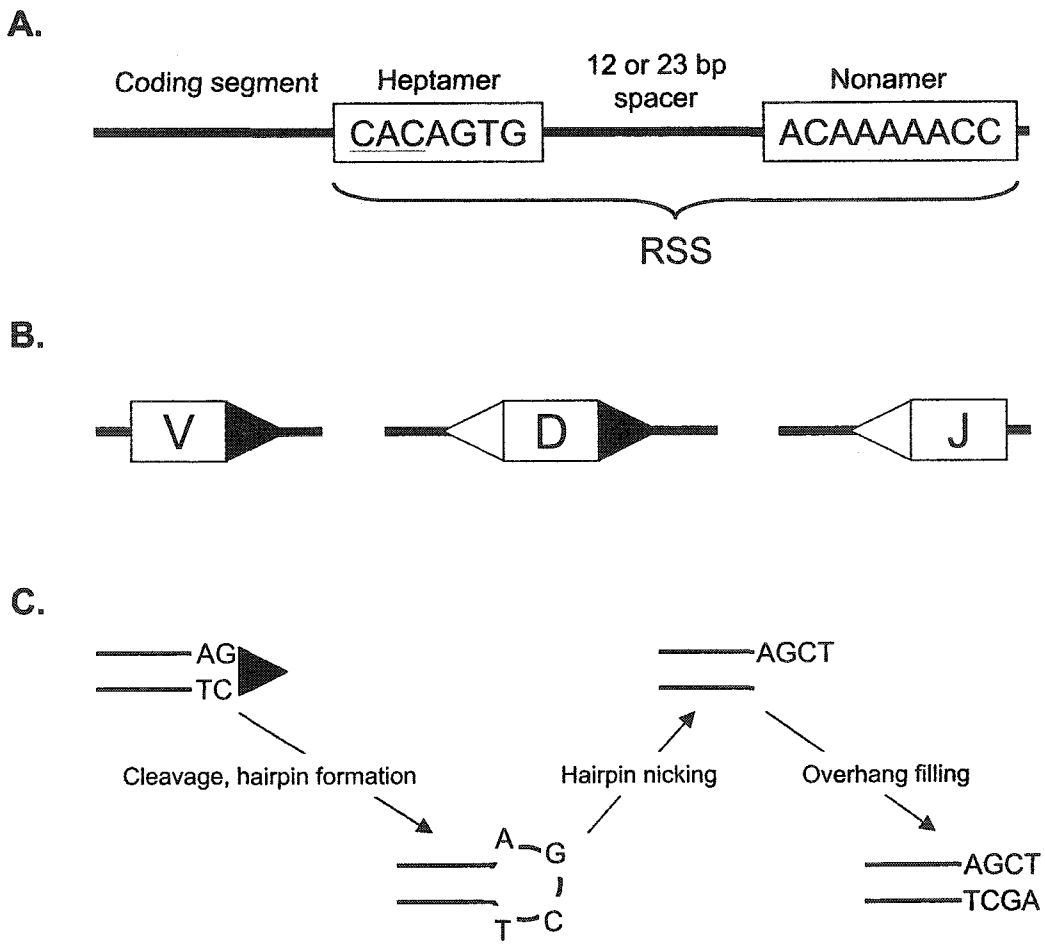
A.
Coding segment
Heptamer
12 or 23 bp spacer
Nonamer
CACAGTG
ACAAAAACC
RSS
B.
V
D
J
C.
AG
TC
AGCT
Cleavage, hairpin formation
Hairpin nicking
Overhang filling
A
G
T
C
AGCT
TCGA

**Figure 2.1**

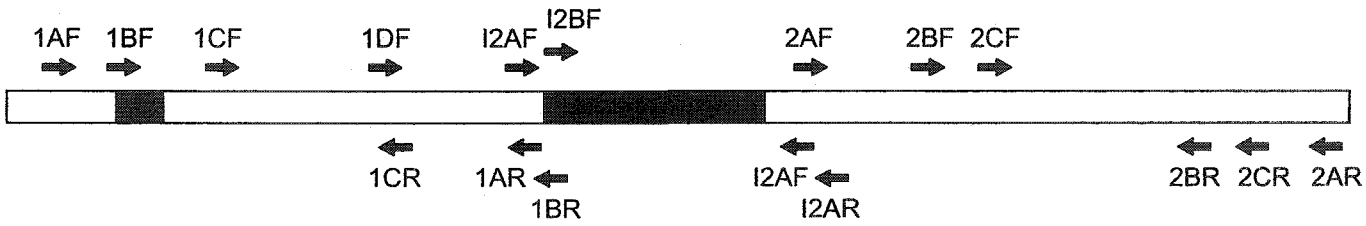
1AF
1BF
1CF
1DF
I2AF
I2BF
2AF
2BF
2CF
1CR
1AR
1BR
I2AF
I2AR
2BR
2CR
2AR

**Figure 2.2**

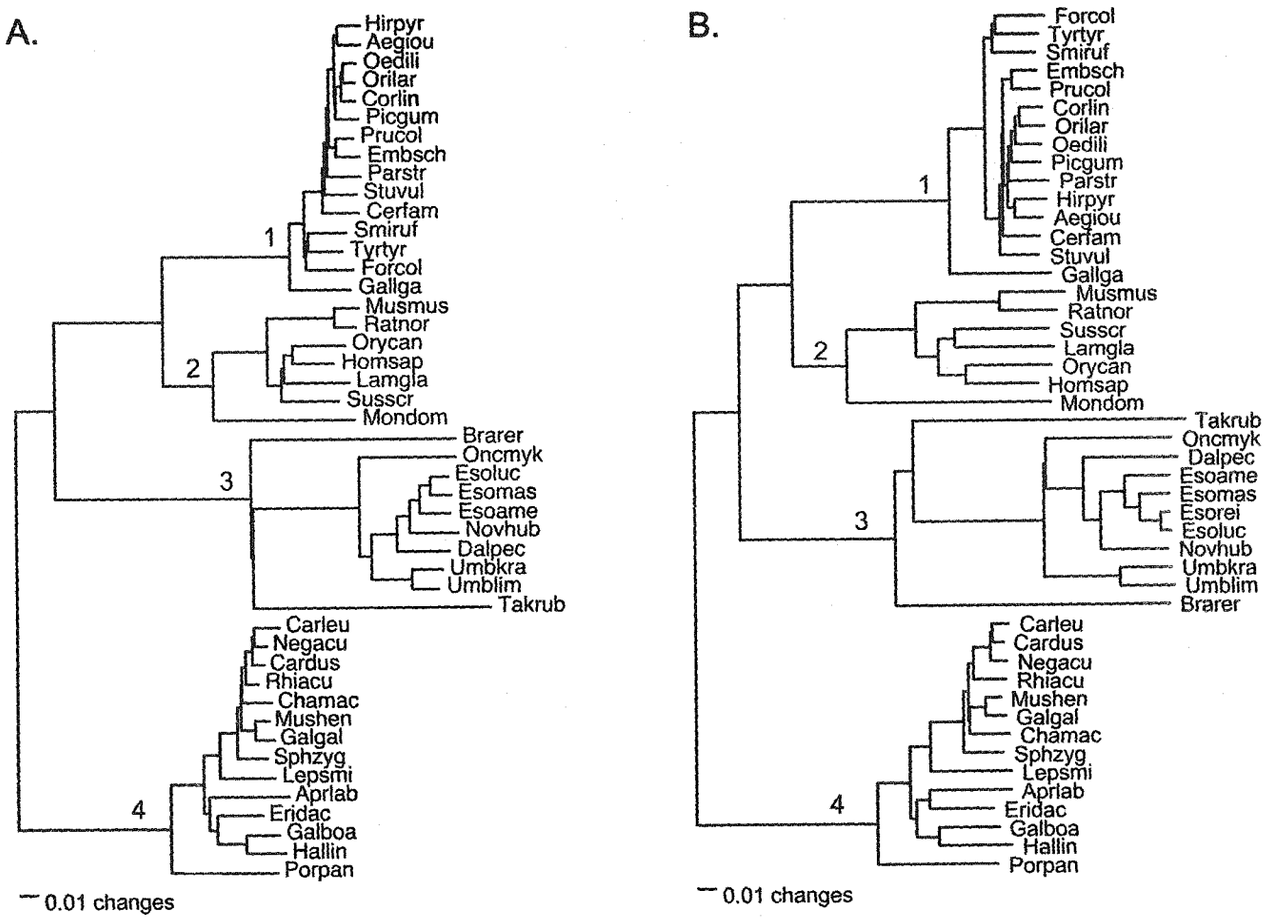
A.
Hirpyr
Aegiou
Oedili
Orilar
Corlin
Picgum
Prucol
Embsch
Parstr
Stuvul
Cerfam
Smiruf
Tyrtyr
Forcol
Gallga
Musmus
Ratnor
Orycan
Homsap
Lamgla
Susscr
Mondom
Brarer
Oncmyk
Esoluc
Esomas
Esoame
Novhub
Dalpec
Umbkra
Umblim
Takrub
Carleu
Negacu
Cardus
Rhiacu
Chamac
Mushen
Galgal
Sphzyg
Lepsmi
Aprlab
Eridac
Galboa
Hallin
Porpan
1
2
3
4
0.01 changes
B.
Forcol
Tyrtyr
Smiruf
Embsch
Prucol
Corlin
Orilar
Oedili
Picgum
Parstr
Hirpyr
Aegiou
Cerfam
Stuvul
Gallga
Musmus
Ratnor
Susscr
Lamgla
Orycan
Homsap
Mondom
Takrub
Oncmyk
Dalpec
Esoame
Esomas
Esorei
Esoluc
Novhub
Umbkra
Umblim
Brarer
Carleu
Cardus
Negacu
Rhiacu
Mushen
Galgal
Chamac
Sphzyg
Lepsmi
Aprlab
Eridac
Galboa
Hallin
Porpan
1
2
3
4
0.01 changes

**Figure 2.3**

**A.**

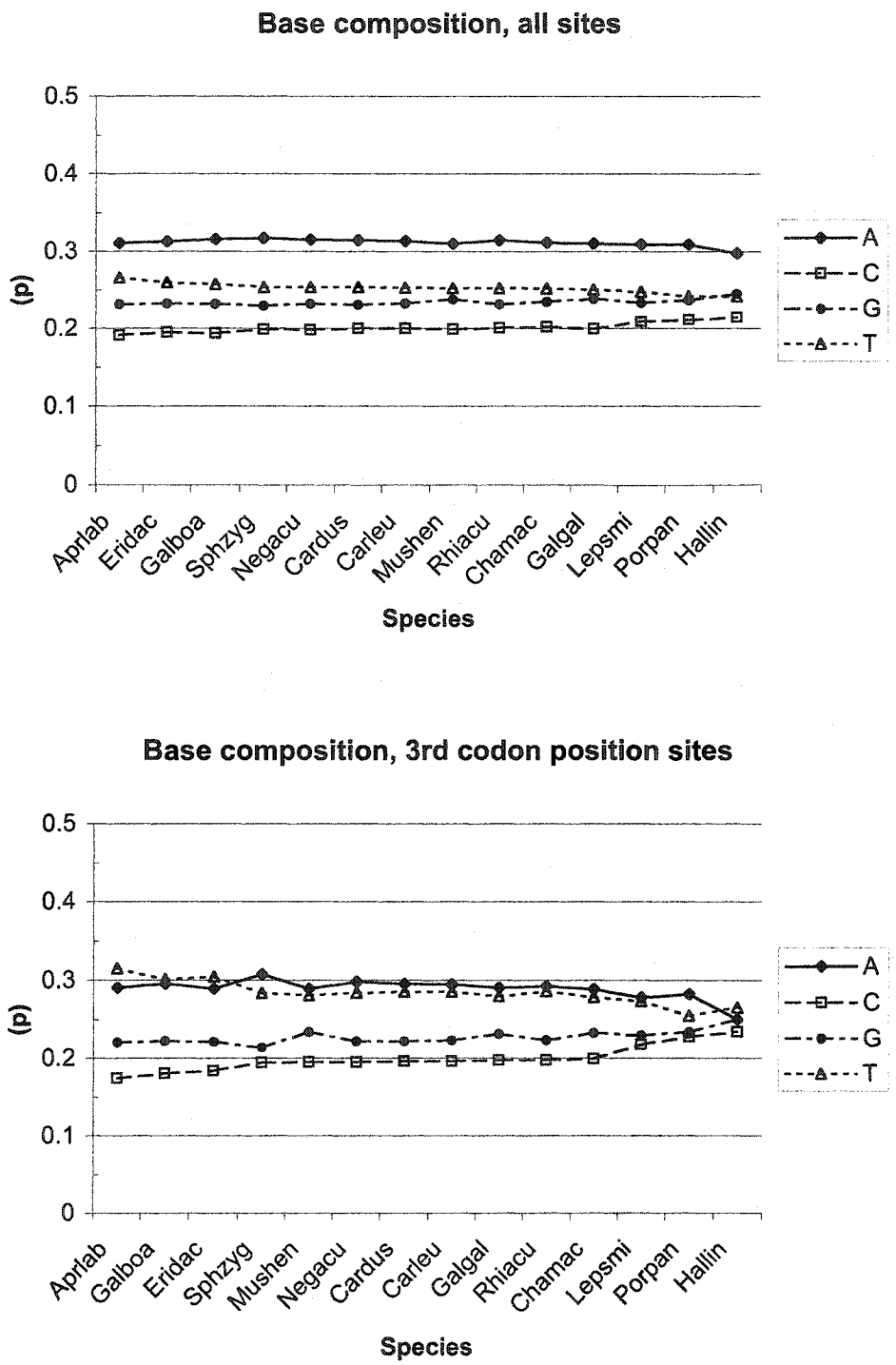
Base composition, all sites
0.5
0.4
0.3
0.2
0.1
0
(p)
A
C
G
T
Aprlab
Eridac
Galboa
Sphzyg
Negacu
Cardus
Carleu
Mushen
Rhiacu
Chamac
Galgal
Lepsmi
Porpan
Hallin
Species
Base composition, 3rd codon position sites
0.5
0.4
0.3
0.2
0.1
0
(p)
A
C
G
T
Aprlab
Galboa
Eridac
Sphzyg
Mushen
Negacu
Cardus
Carleu
Galgal
Rhiacu
Chamac
Lepsmi
Porpan
Hallin
Species

**Figure 2.4**

**B.**

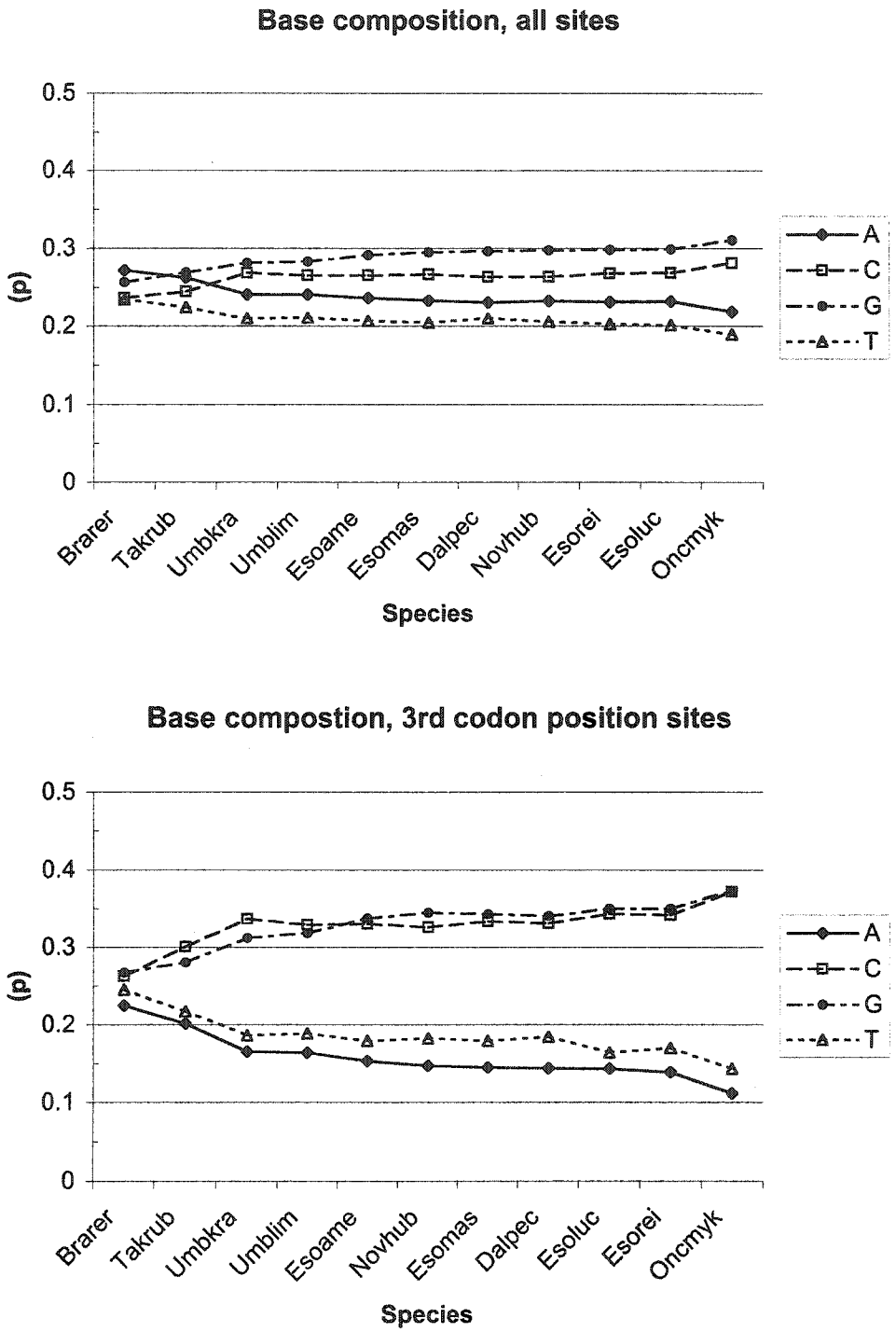


**Figure 2.4** (continued)

C.

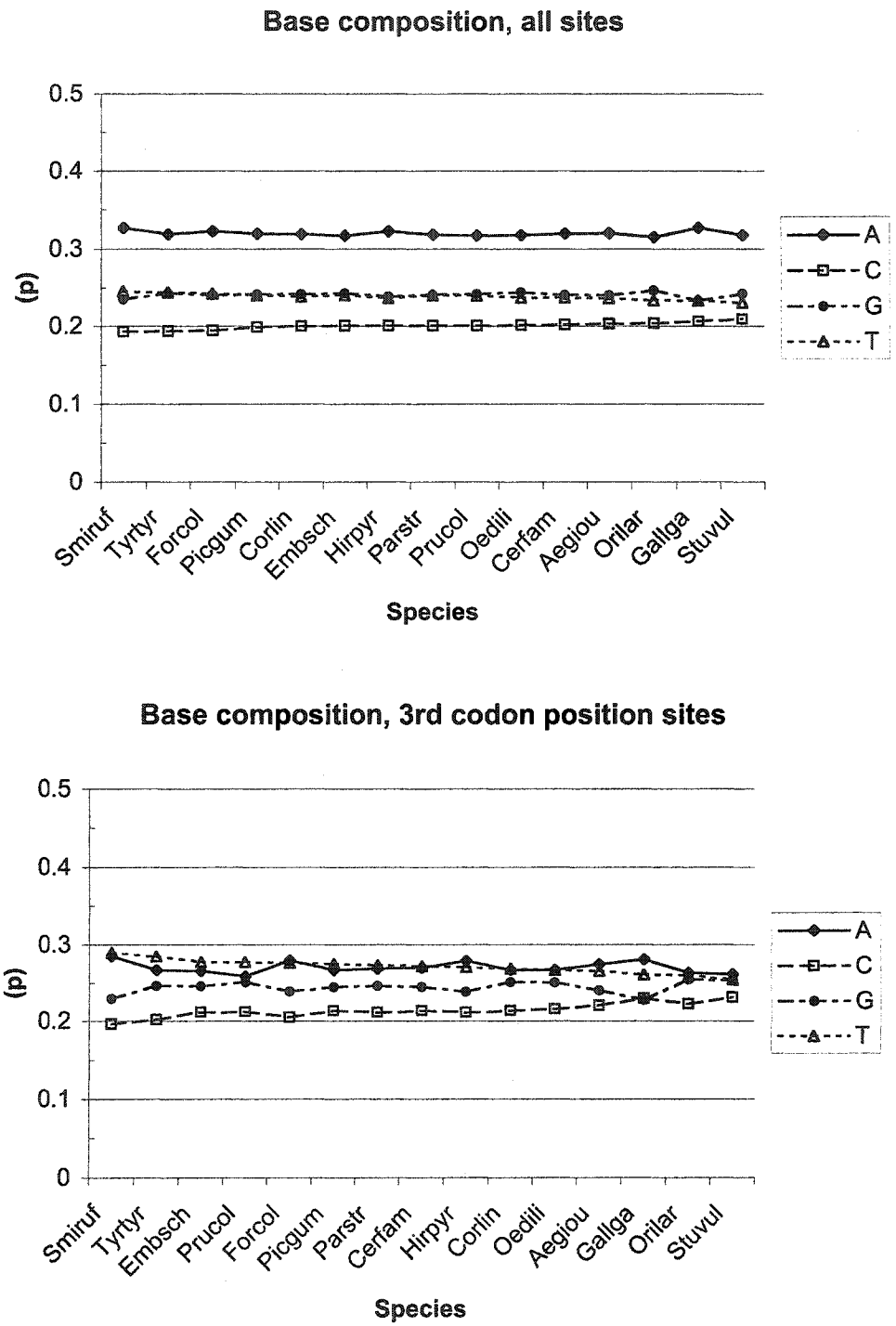


**Figure 2.4** (continued)

D.

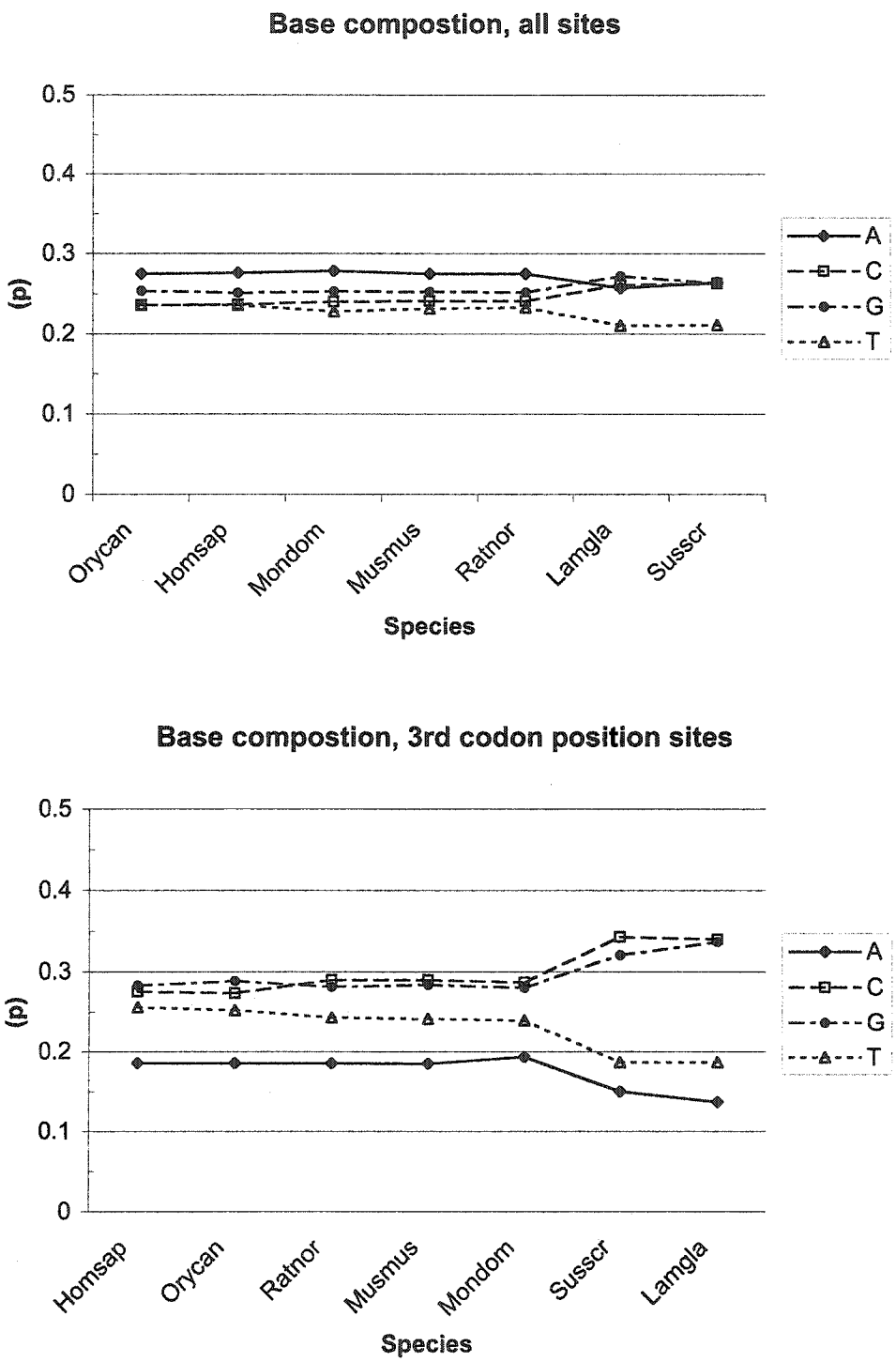


**Figure 2.4** (continued)

**E.**

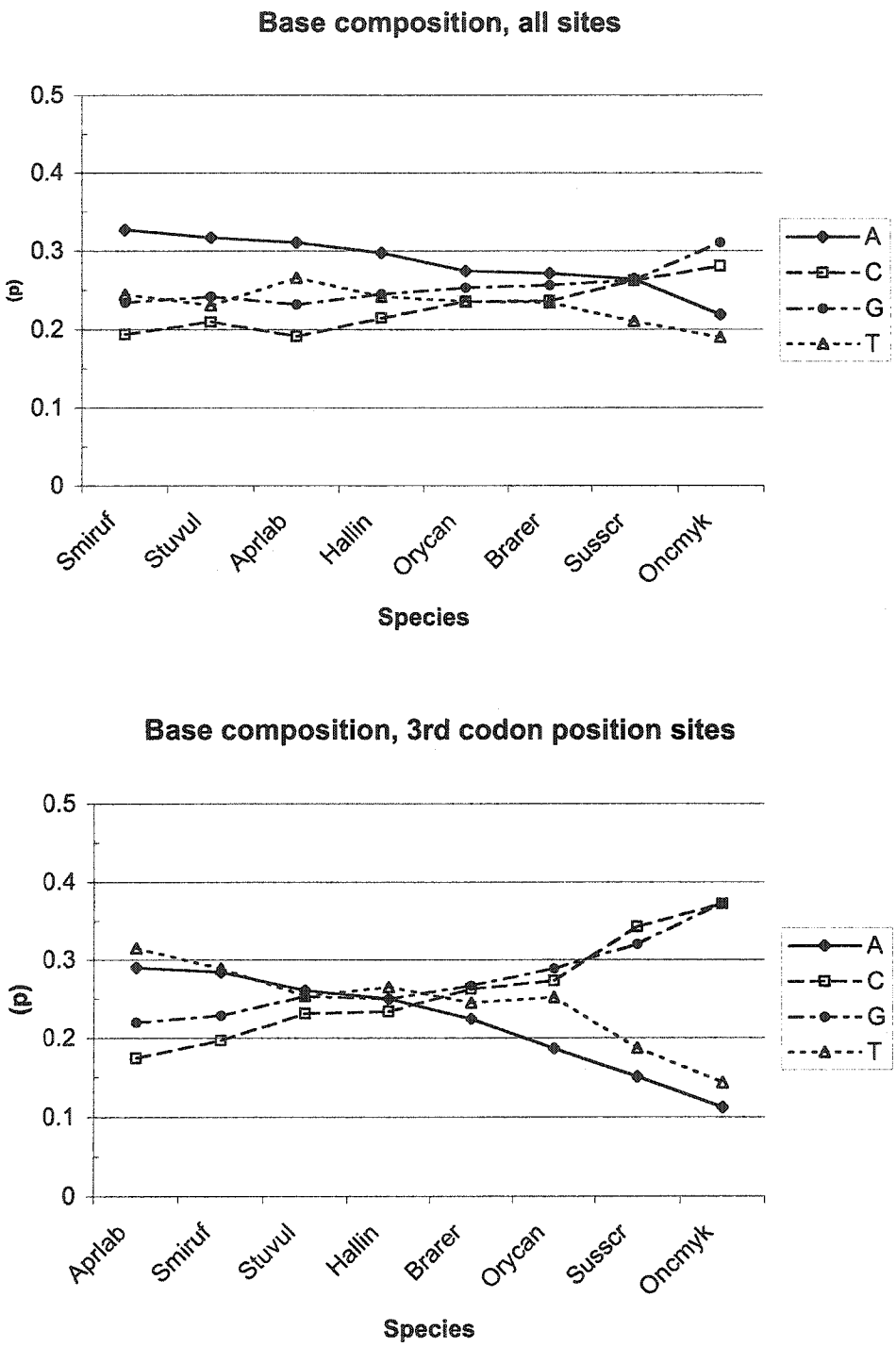


**Figure 2.4** (continued)

F.

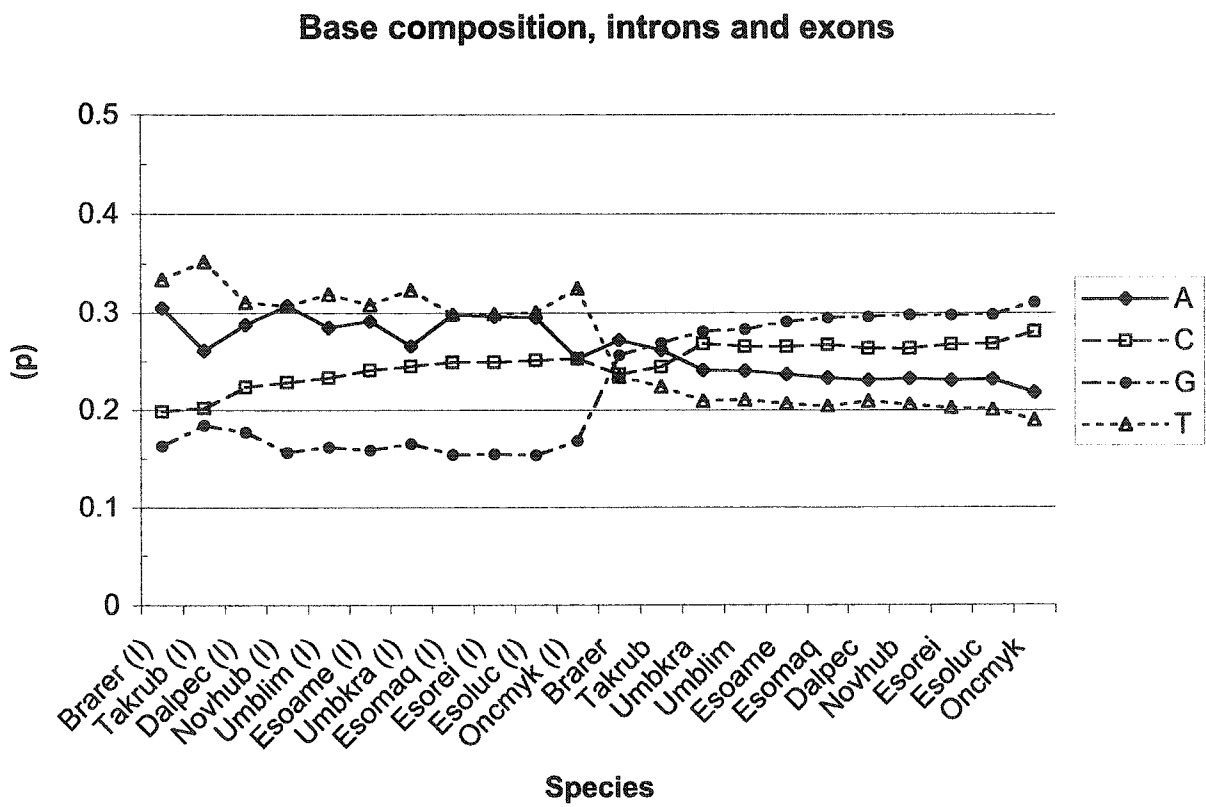


**Figure 2.4** (continued)

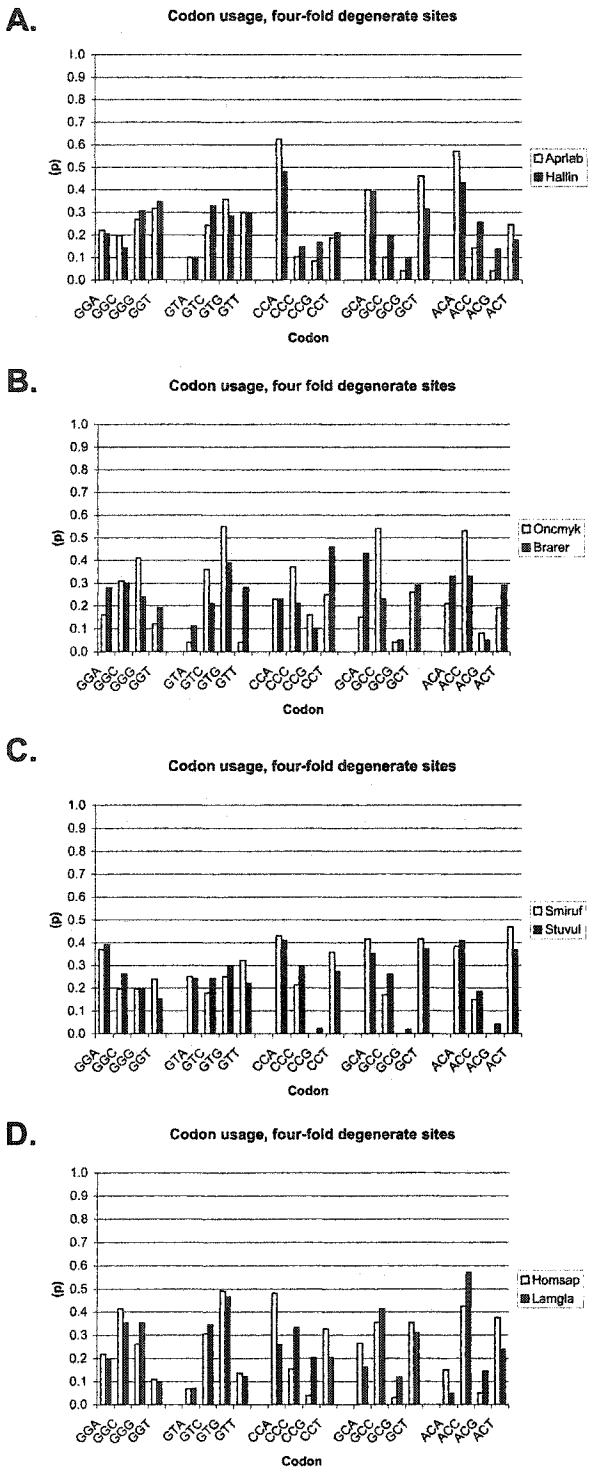
A.
Codon usage, four-fold degenerate sites
(p)
Codon
Aprlab
Hallin
B.
Codon usage, four fold degenerate sites
(p)
Codon
Oncmyk
Brarer
C.
Codon usage, four-fold degenerate sites
(p)
Codon
Smiruf
Stuvul
D.
Codon usage, four-fold degenerate sites
(p)
Codon
Homsap
Lamgla
GGA GGC GGG GGT GTA GTC GTG GTT CCA CCC CCG CCT GCA GCC GCG GCT ACA ACC ACG ACT

**Figure 2.5**

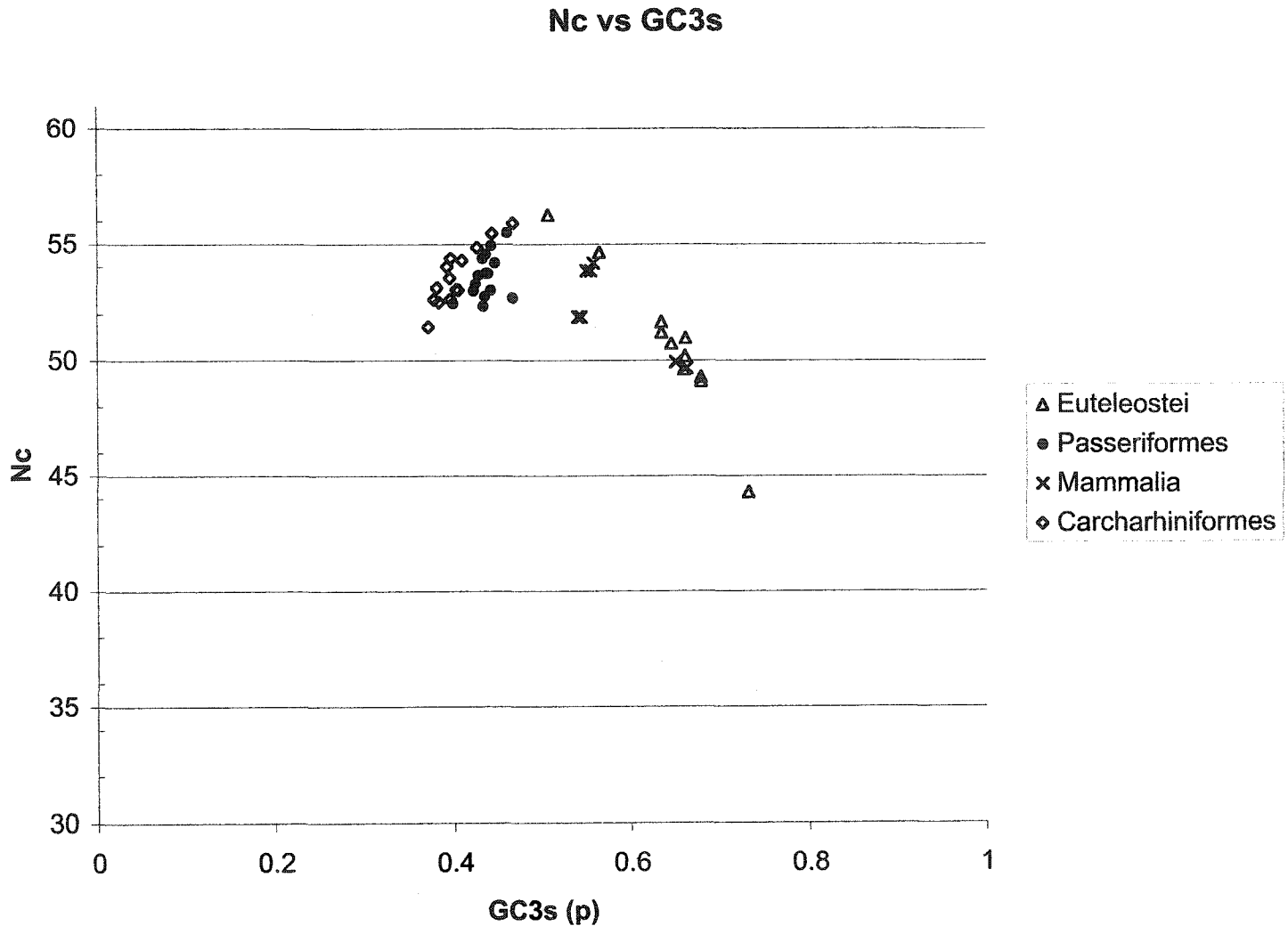
Nc vs GC3s
60
55
50
45
40
35
30
Nc
0
0.2
0.4
0.6
0.8
1
GC3s (p)
Euteleostei
Passeriformes
Mammalia
Carcharhiniformes

**Figure 2.6**

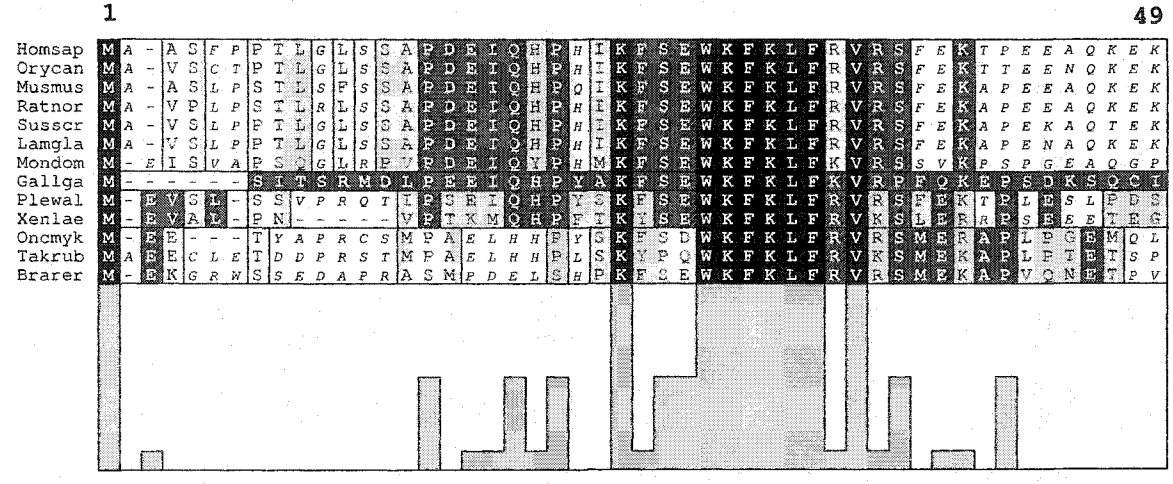
1
49
Homsap
Orycan
Musmus
Ratnor
Susscr
Lamgla
Mondom
Gallga
Plewal
Xenlae
Oncmyk
Takrub
Brarer

**Figure 2.7**

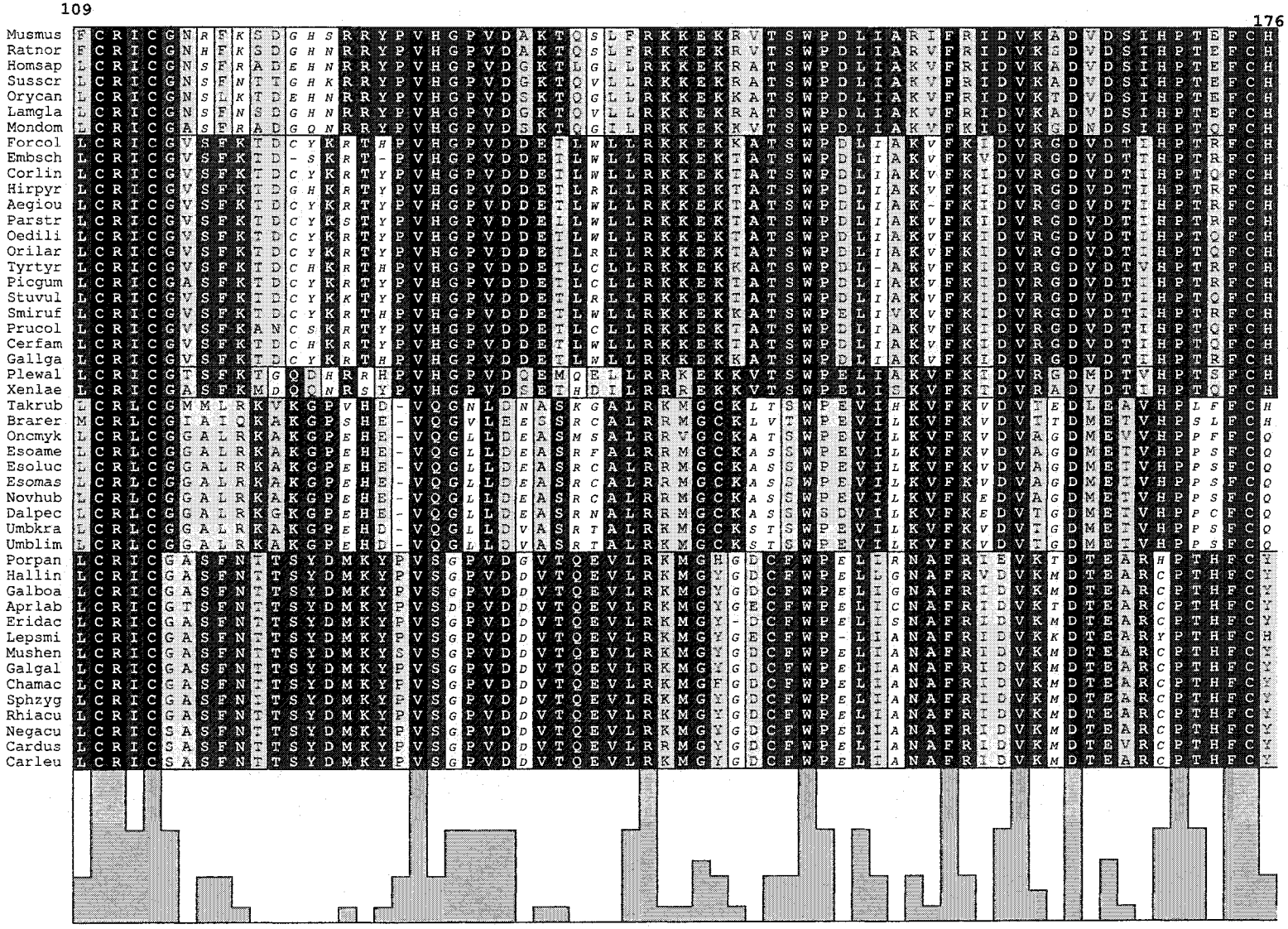
109
176
Musmus
Ratnor
Homsap
SussCr
Orycan
Lamgla
Mondom
Forcol
Embsch
Corlin
Hirpyr
Aegiou
Parstr
Oedili
Orilar
Tyrtyr
Picgum
Stuvul
Smiruf
Prucol
Cerfam
Gallga
Plewal
Xenlae
Takrub
Brarer
Oncmyk
Esoame
Esoluc
Esomas
Novhub
Dalpec
Umbkra
Umblim
Porpan
Hallin
Galboa
Aprlab
Eridac
Lepsmi
Mushen
Galgal
Chamac
Sphzyg
Rhiacu
Negacu
Cardus
Carleu

Figure 2.8

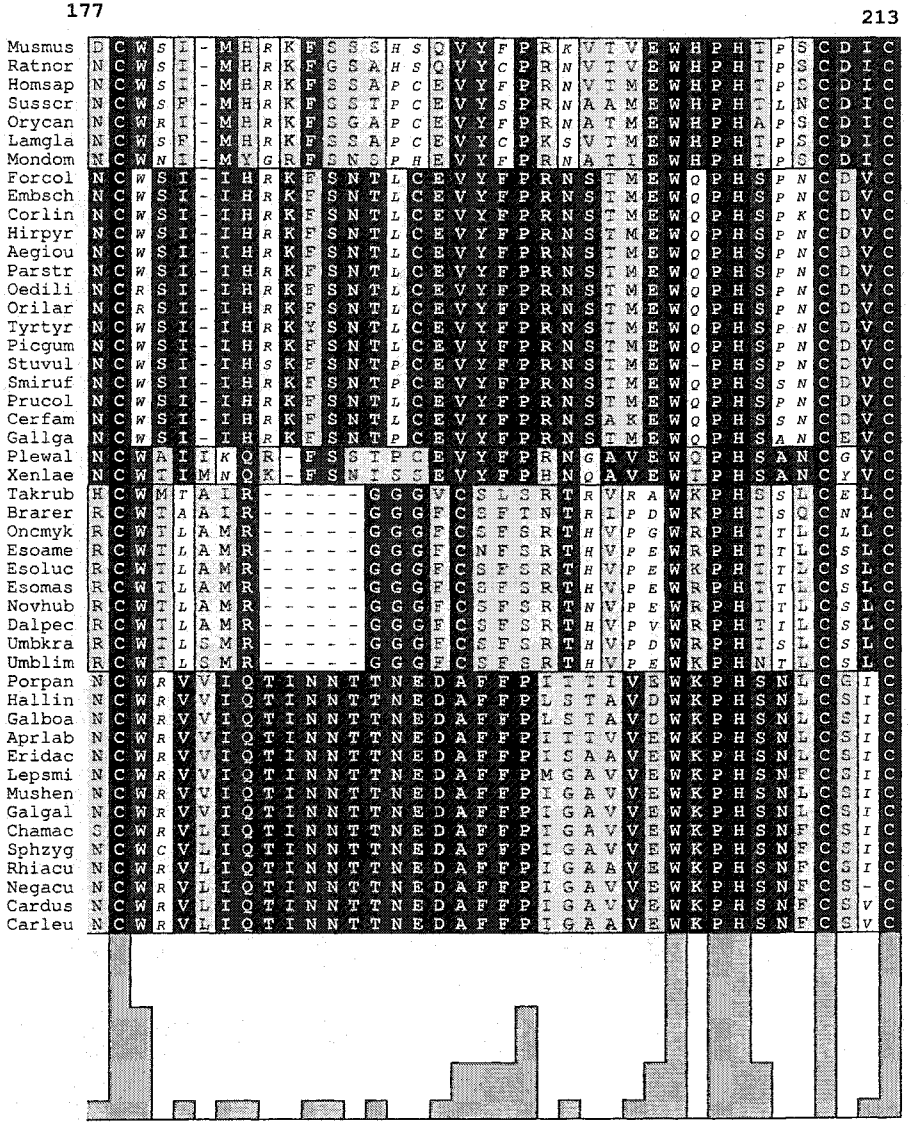


**Figure 2.8** (continued)

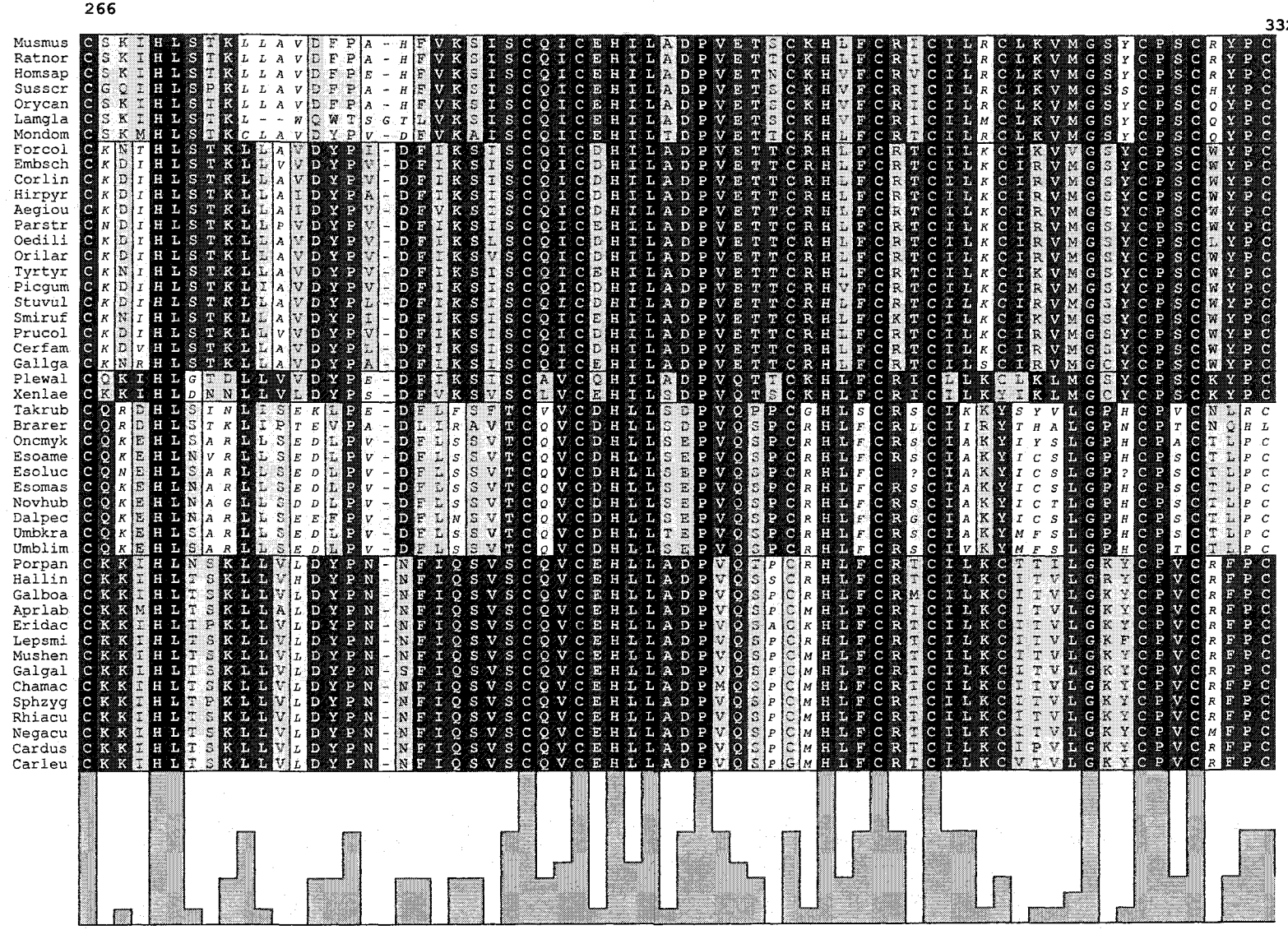
266
332
Musmus
Ratnor
Homsap
Susscr
Orycan
Lamgla
Mondom
Forcol
Embsch
Corlin
Hirpyr
Aegiou
Parstr
Oedili
Orilar
Tyrtyr
Picgum
Stuvul
Smiruf
Prucol
Cerfam
Gallga
Plewal
Xenlae
Takrub
Brarer
Oncmyk
Esoame
Esoluc
Esomas
Novhub
Dalpec
Umbkra
Umblim
Porpan
Hallin
Galboa
Aprlab
Eridac
Lepsmi
Mushen
Galgal
Chamac
Sphzyg
Rhiacu
Negacu
Cardus
Carleu

**Figure 2.9**

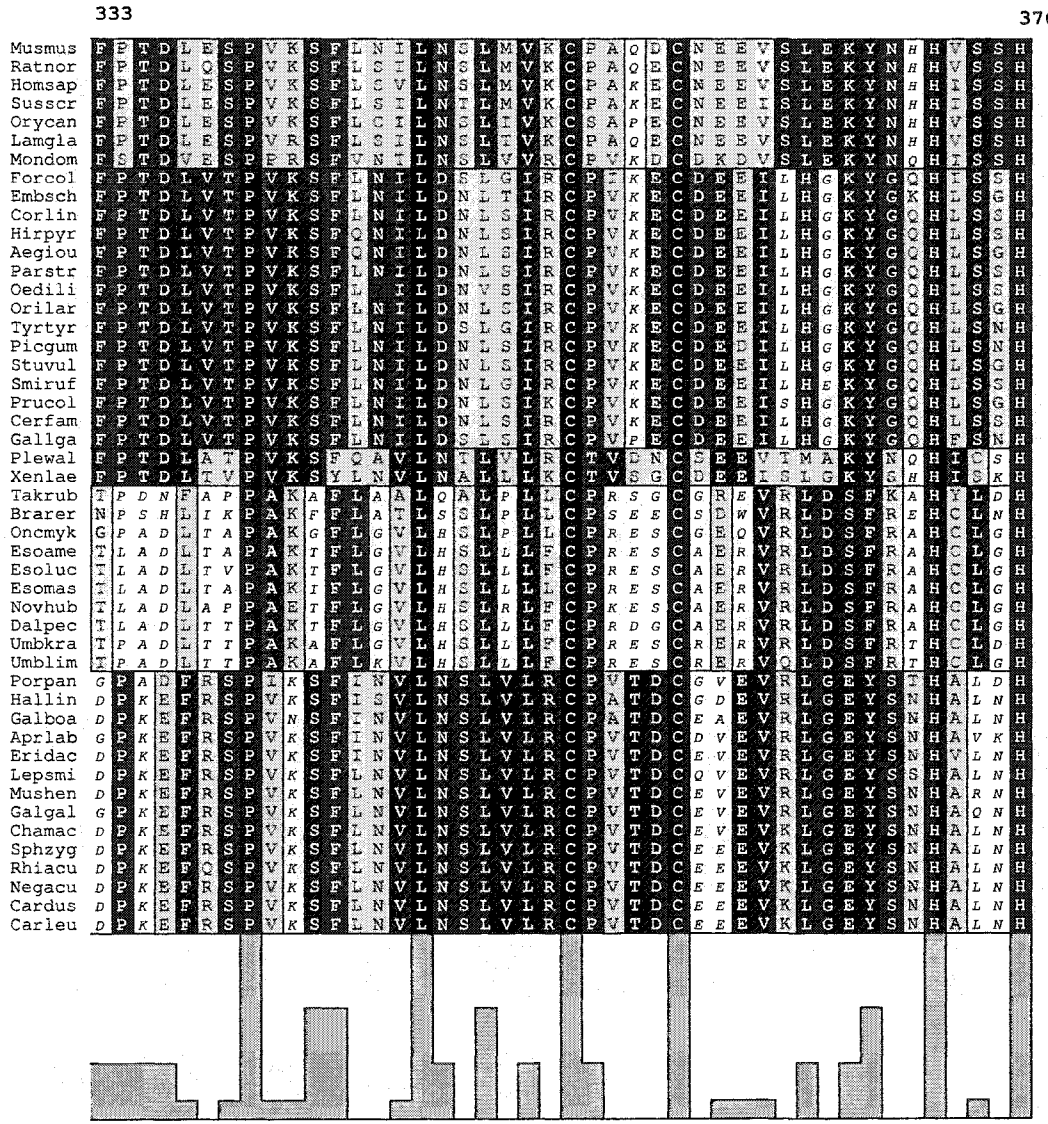


**Figure 2.9** (continued)

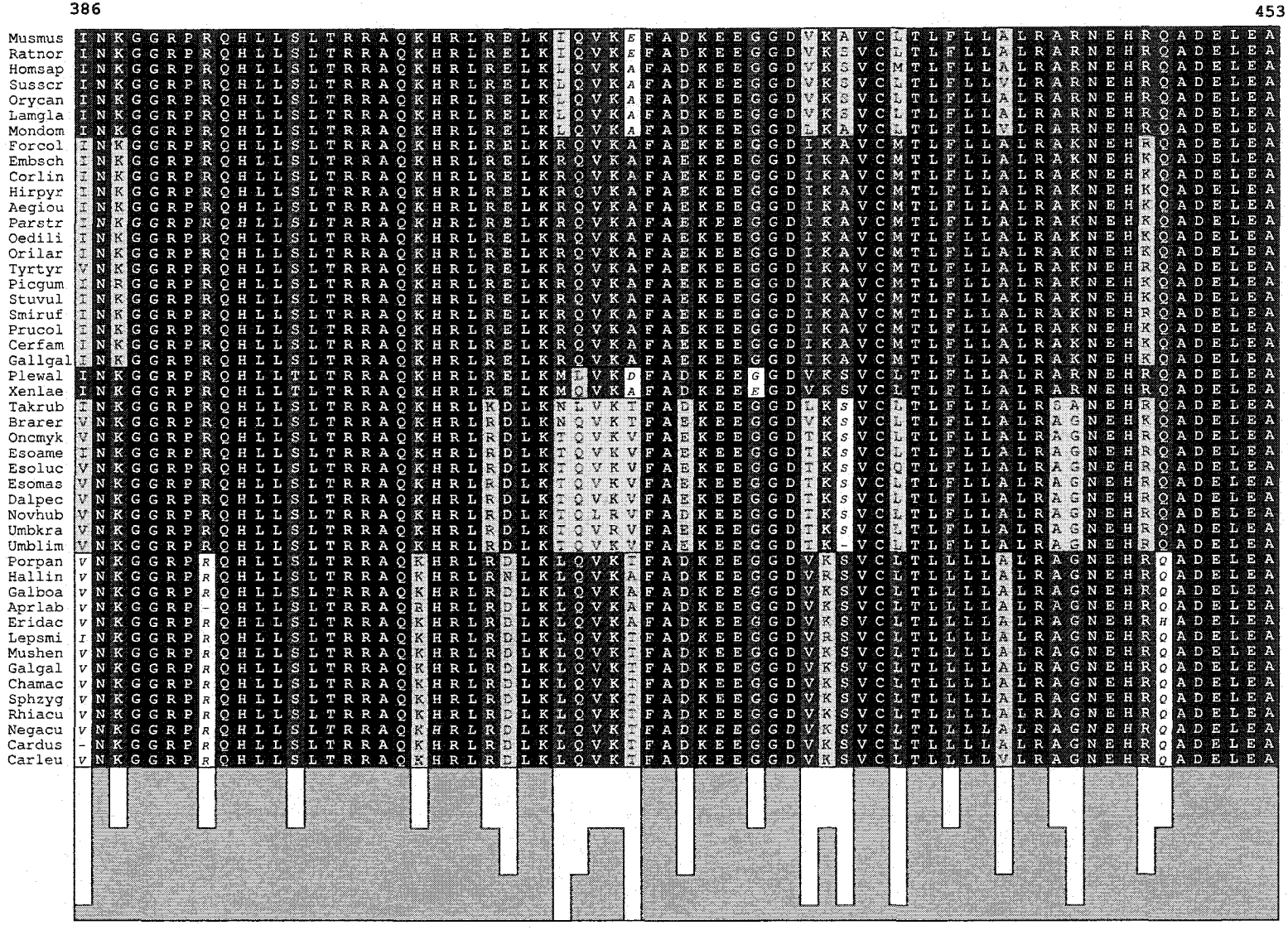
386
453
Musmus
Ratnor
Homsap
Susscr
Orycan
Lamgla
Mondom
Forcol
Embsch
Corlin
Hirpyr
Aegiou
Parstr
Oedili
Orilar
Tyrtyr
Picgum
Stuvul
Smiruf
Prucol
Cerfam
Gallgal
Plewal
Xenlae
Takrub
Brarer
Oncmyk
Esoame
Esoluc
Esomas
Dalpec
Novhub
Umbkra
Umblim
Porpan
Hallin
Galboa
Aprlab
Eridac
Lepsmi
Mushen
Galgal
Chamac
Sphzyg
Rhiacu
Negacu
Cardus
Carleu

**Figure 2.10**

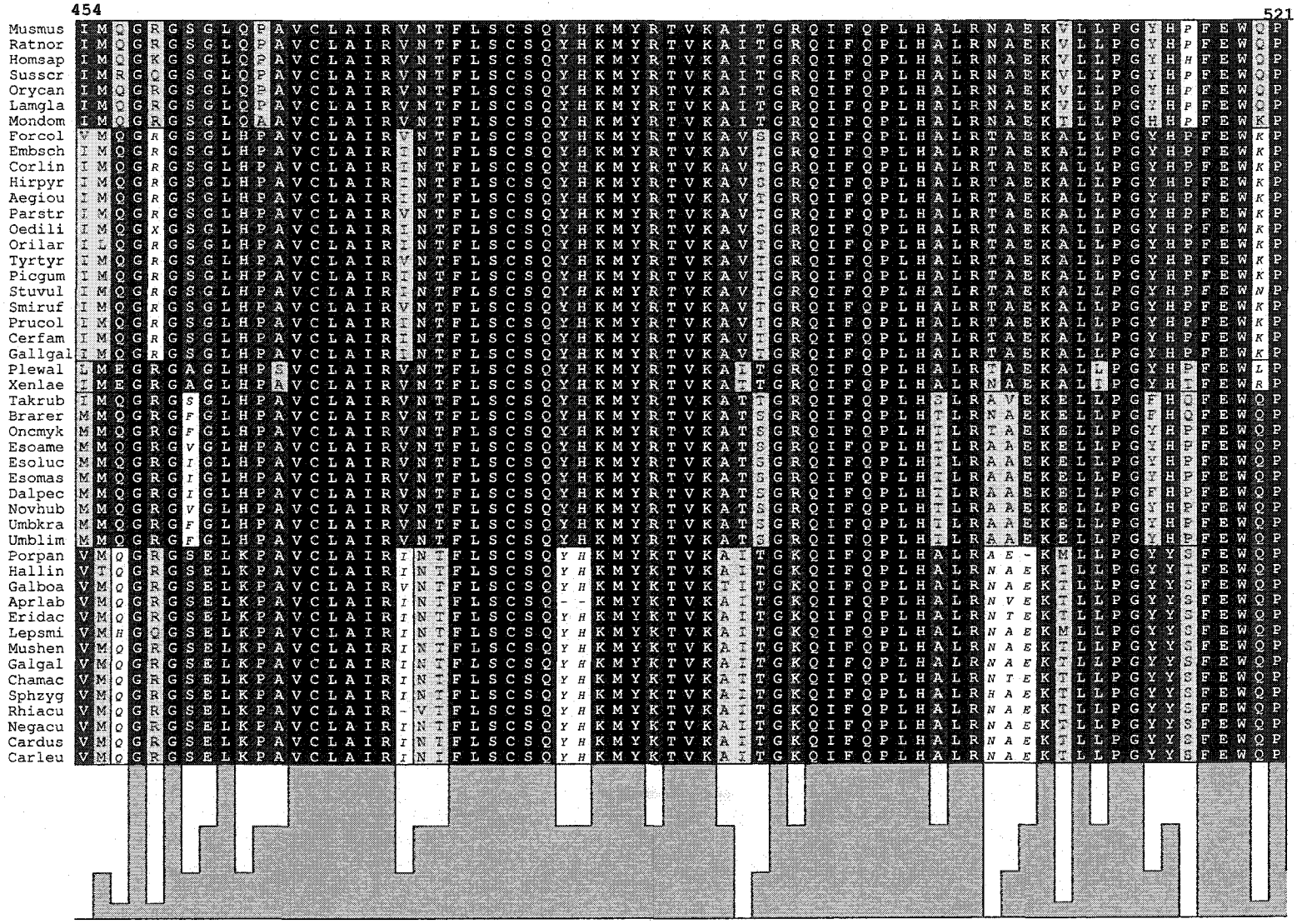


**Figure 2.10** (continued)

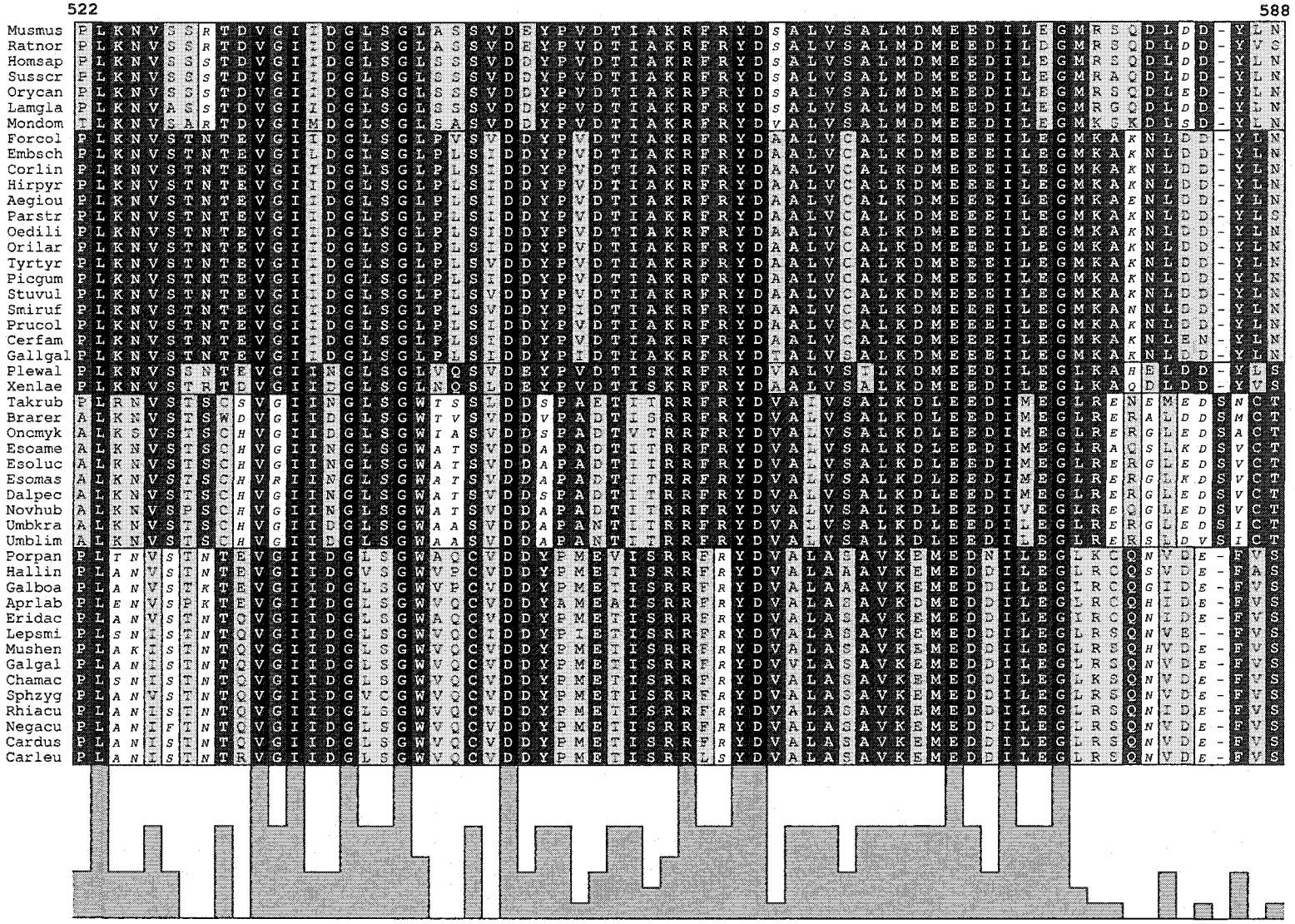


**Figure 2.10** (continued)

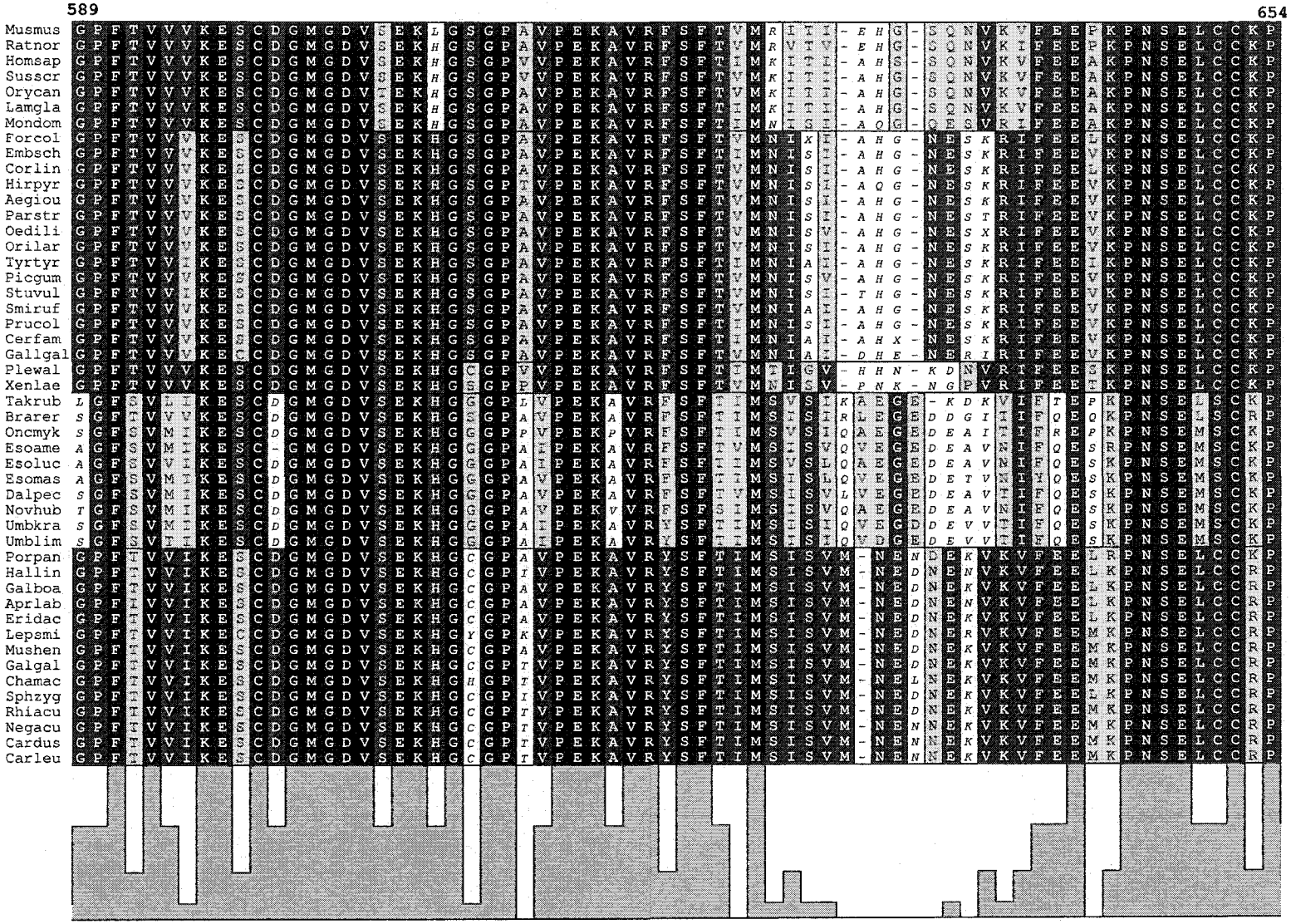


**Figure 2.10** (continued)

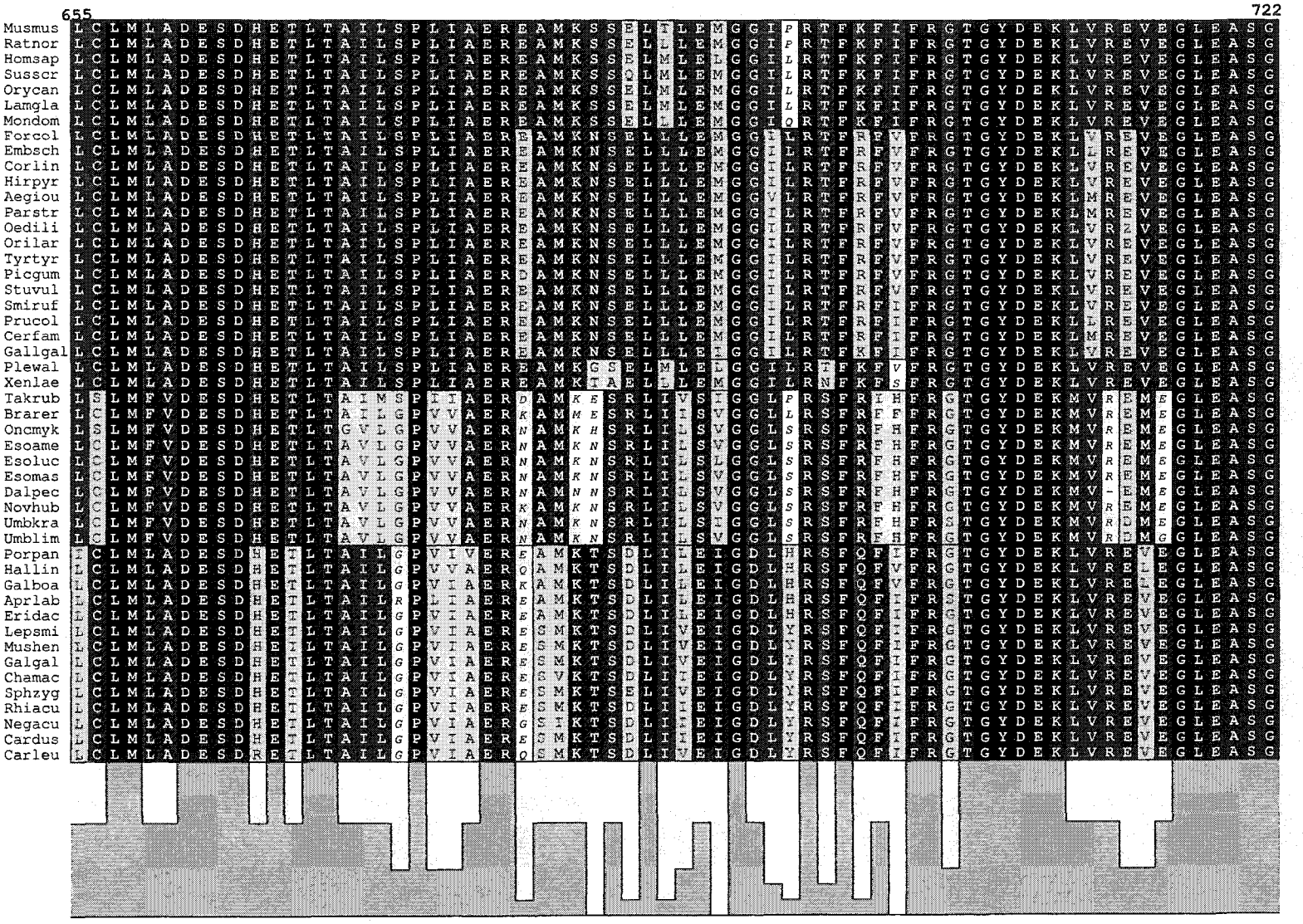


**Figure 2.10** (continued)

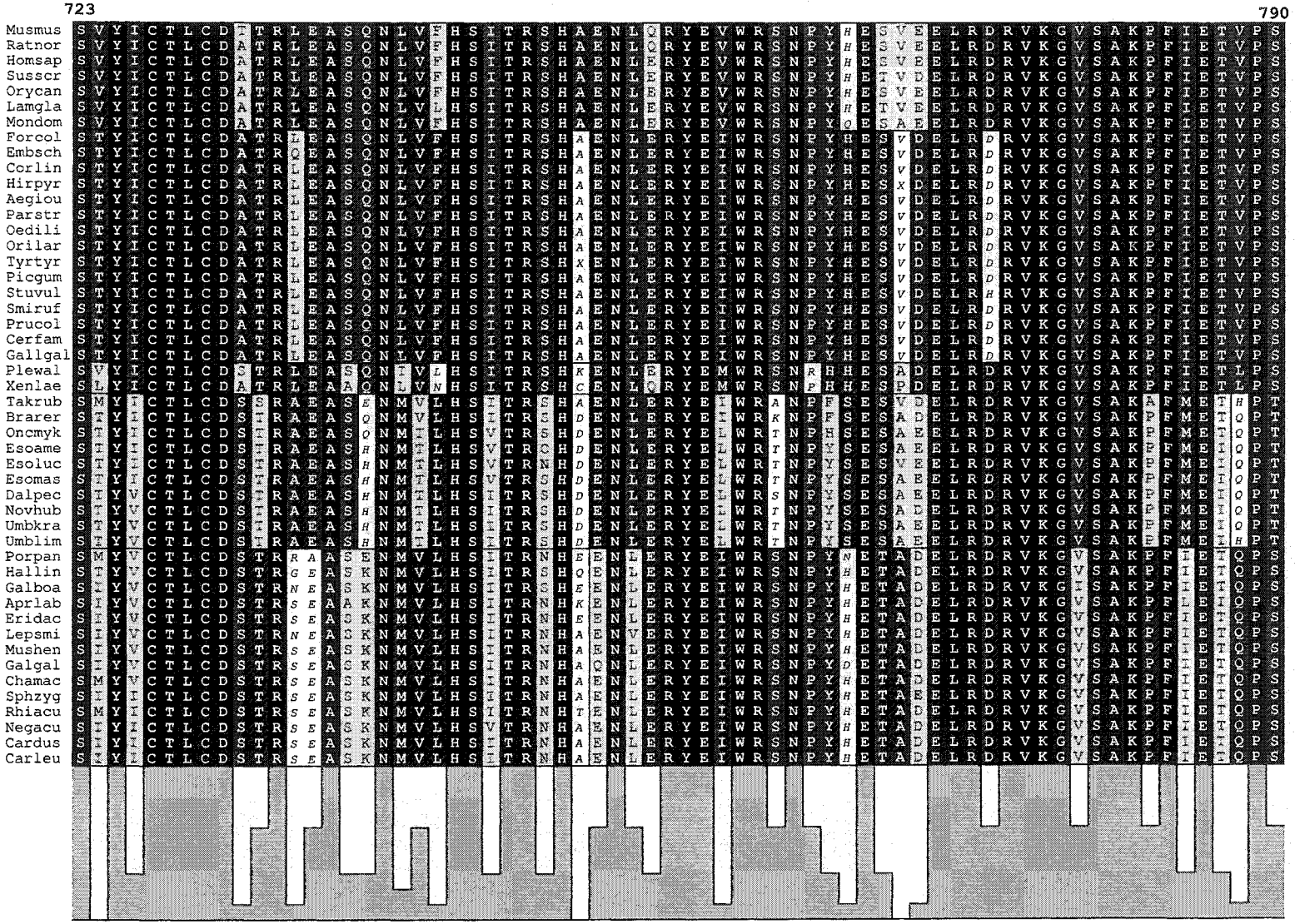


**Figure 2.10** (continued)

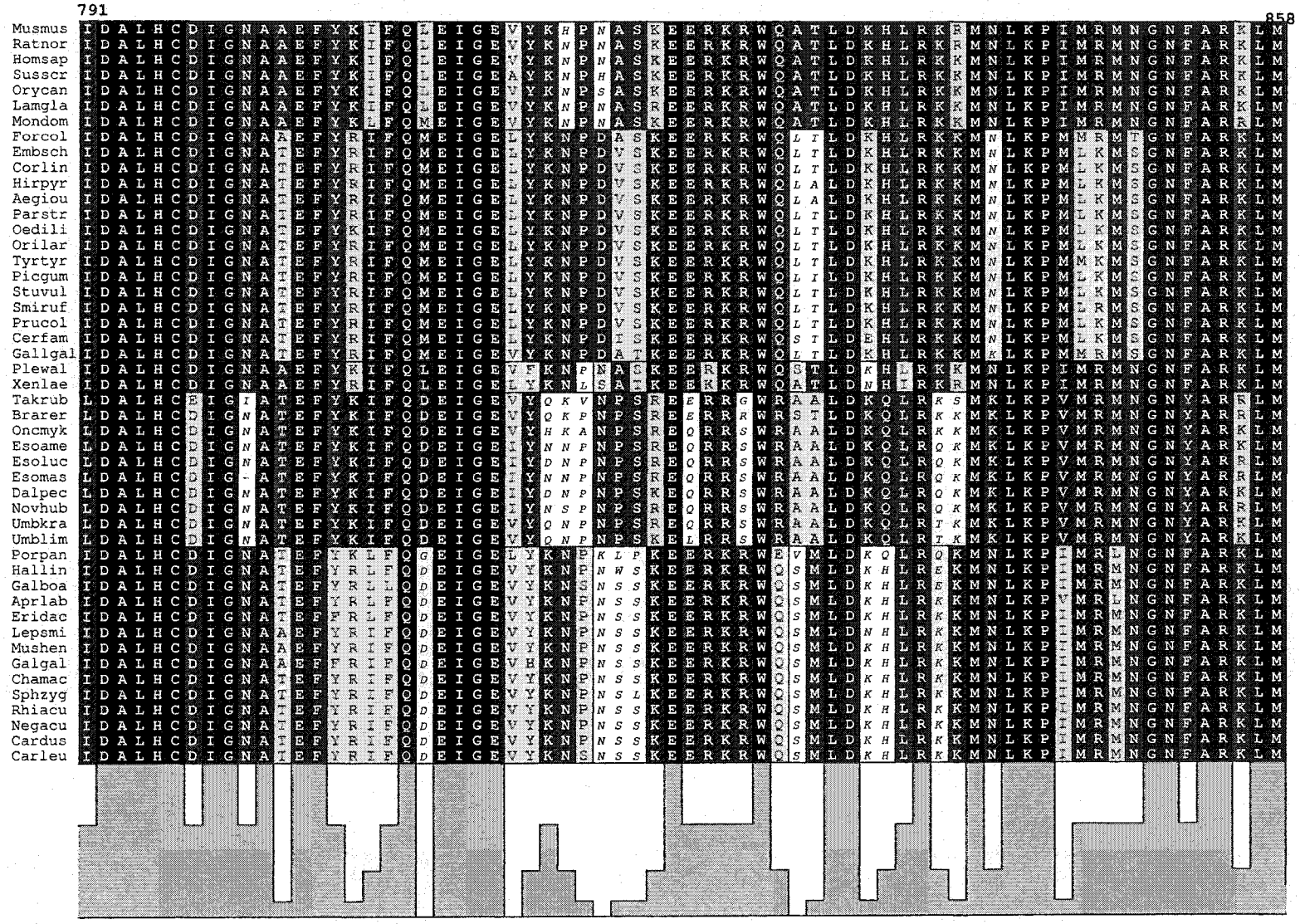


**Figure 2.10** (continued)

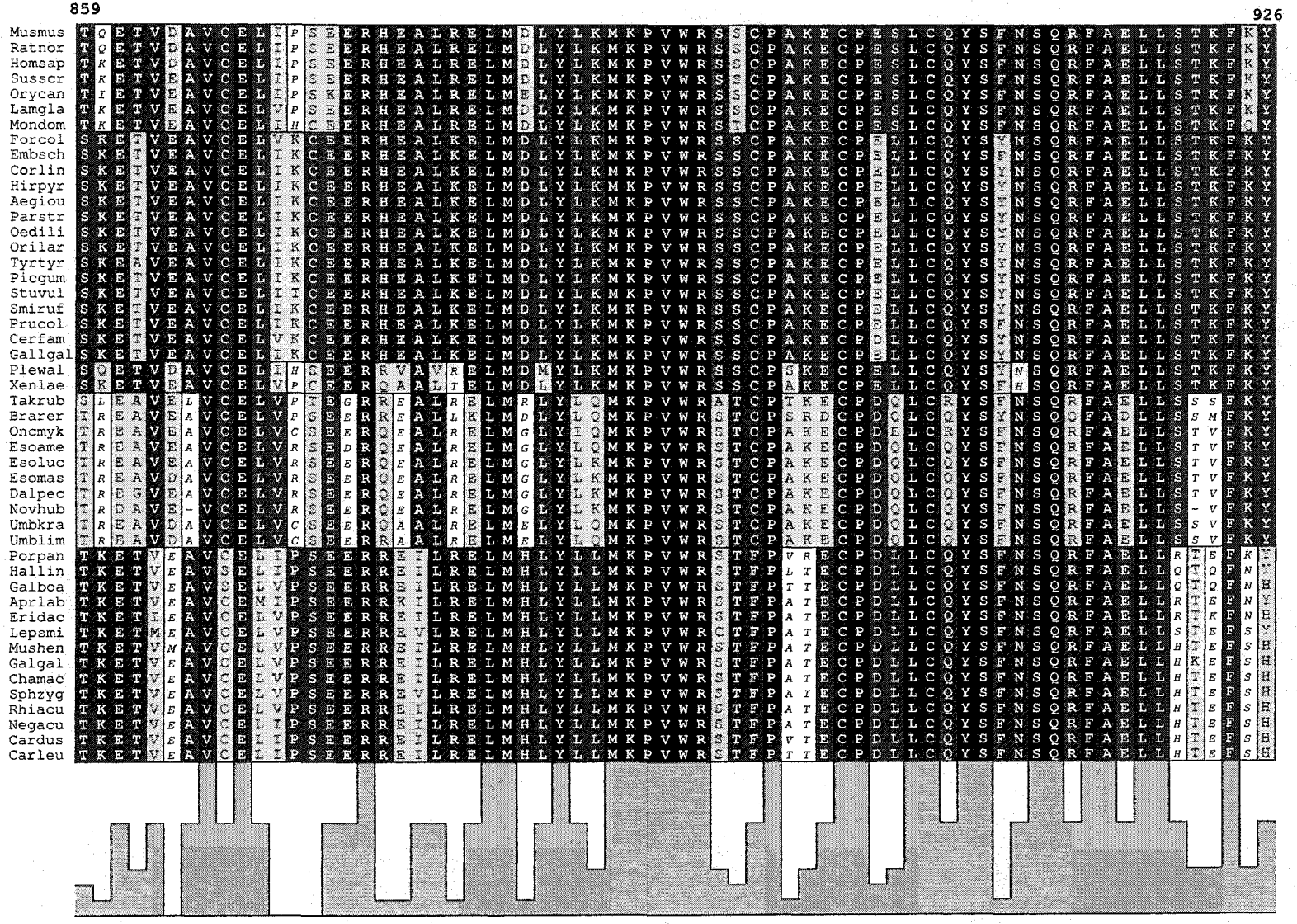


**Figure 2.10** (continued)

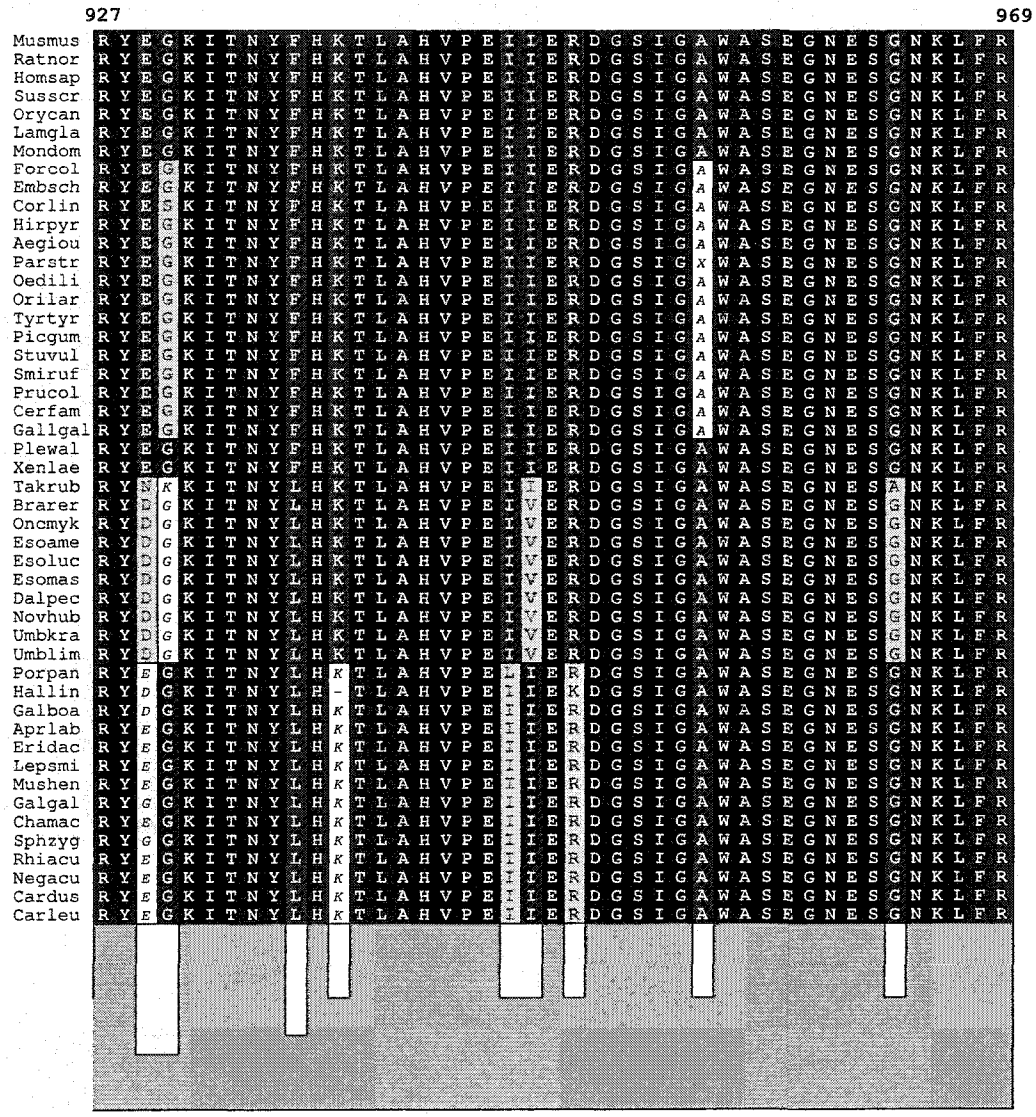


**Figure 2.10** (continued)

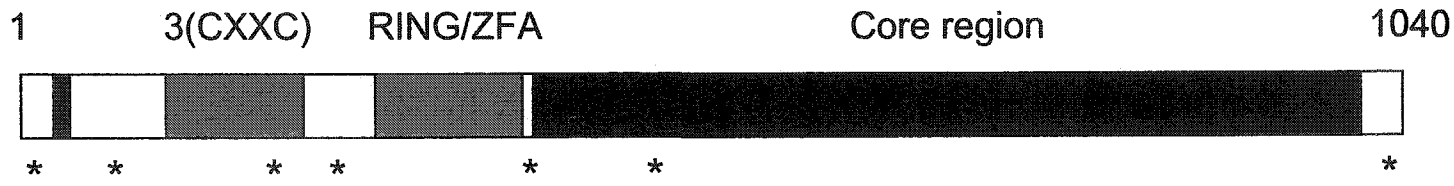


**Figure 2.11**

# CHAPTER 3. ESOCIFORM RELATIONSHIPS

A paper submitted to *Copeia*

Juan Andrés López, Wei-Jen Chen and Guillermo Orti

**Abstract**

Despite numerous studies that aimed at resolving relationships among basal euteleost lineages, many aspects of their phylogeny remain unresolved. Among euteleost lineages, the Esociformes have proven particularly difficult to place. Among esociform taxa, Nelson's (1972) hypothesis of relationships has been generally accepted but recently available evidence from molecular sequences contradicted that view. We have assembled a data set of DNA sequences from the mitochondrial and nuclear genomes to test existing hypotheses of esociform relationships. This data set includes representatives from all extant esociform lineages and a wide diversity of potential outgroups. We also conducted a critical review of the morphological information that supports currently held views of esociform inter- and intra-ordinal relationships. Our review revealed several problems with character state coding and interpretation of character states. In contrast, the molecular evidence, particularly the nuclear sequences, produced strong support for a sister group relationship between esociformes and salmonoids. The DNA sequences also offer strong corroboration of the hypothesis of esociform intraordinal relationships of López et al. (2000) and for the monophyly of the subgenera *Esox* and *Kenoza* of *Esox*. In addition to the conclusions regarding esociform relationships, the molecular evidence we present offers support for the monophyly of the Osmeridae and for a sister group relationship between the Retropinnidae and the Osmeroidei (Osmeridae + Salangidae + Plecoglossidae).

## Introduction

The basal euteleost lineages include those fish groups that Greenwood et al. (1966) placed in the Protacanthopterygii, erected to collect the most primitive members of their Division III of teleost fishes. Despite repeated study of the phylogenetic relationships of the protacanthopterygian lineages a robust hypothesis remains elusive and alternative hypotheses abound. As a consequence of this diversity of opinion, protacanthopterygian membership has been modified extensively since its inception; at one point being reduced to the Salmoniformes (Rosen, 1973) and thus becoming a redundant taxonomic concept.

Although ichthyologists have not yet arrived at a generally accepted reconstruction of basal euteleost relationships there are some common features among the hypotheses advanced to date. A basal euteleost lineage that has been consistently difficult to place is the Esociformes (pikes, pickerels and mudminnows). So far, searches for suites of derived character states that may point to the esociform sister group have yielded scarce evidence resulting in weakly supported hypotheses. It seems that the gross morphology of esociforms is largely characterized by autapomorphy and euteleost plesiomorphy. Esociformes have been proposed to be: (1) sister taxon of all other euteleosts (Fink and Weitzman, 1982), (2) sister taxon of all other protacanthopterygians (with *Lepidogalaxias*; Rosen, 1974), (3) sister taxon of the salmonoids (Fink, 1984) and (4) sister taxon of the neoteleosts (Parenti, 1986; Johnson and Patterson, 1996).

Johnson and Patterson's (1996) study is the most recent, extensive and comprehensive cladistic analysis of basal euteleosts relationships available. Their results support the following conclusions regarding esociform relationships: (1) tentatively, a sister group relationship of Esociformes and Neoteleostei (Fig 3.1A) and (2) agreement with the hypothesis of relationships among esociform genera of Nelson (1972; Fig 3.1B).

*The molecular evidence*

López et al. (2000) reported evidence based on DNA sequences of two mitochondrial genes that supported an arrangement of esociform genera incompatible with Nelson's (1972) hypothesis and that suggested a close relationship between Esociformes and Salmoniformes. In López et al.'s (2000) study the monophyly of the Umbridae was strongly rejected. *Umbra* was proposed to be the most basal esociform genus and, *Novumbra* and *Esox* were supported as sister genera (Fig. 1C). That study suffered from the following weaknesses: (1) it was based on an incomplete sample of species of *Esox*, (2) a very limited set of alternative outgroups were sampled (López et al., 2000; Material Examined – p. 429), (3) the sequence data were all derived from the mitochondrial genome and (4) the strongest evidence in support of their conclusions came from the 16S sequences, which show levels of divergence among esociformes large enough to suggest potential problems with the alignment of some extremely variable loop regions.

Other evidence that suggests that Nelson's (1972) hypothesis of esociform phylogeny does not accurately reflect the evolutionary history of the group comes from a number of studies of the karyotypes and other cytological characteristics of esociforms. In the most recent of these studies, Crossman and Ráb (2001) surmised that the karyological characteristics of esociforms point to a close relationship between *Dallia* and *Novumbra* to the exclusion of *Umbra* and an uncertain placement of *Esox* in relation to these two lineages. Unfortunately, the methods employed so far in the study of esociform karyology produce observations that are difficult to frame in the context of a phylogenetic analysis with the requisite assertions of homology of chromosome elements. Thus, although intriguing, the karyological data currently available are not sufficient to reassess esociform relationships.

There have been a number of molecular studies that support a close relationship between esociformes and salmonoids. Most of these were based on incomplete taxonomic samples that did not allow firm conclusions regarding esociform relationships (see summary in Zaragueta-Bagils et al., 2002). The most thorough molecular study of protacanthopterygian relationships is that of Ishiguro et al. (in press), using whole mitochondrial genome sequences. Their analysis support (82% bootstrap) a sister group relationship between the two esociforms in their study and the salmonoids.

*The fossil record*

The paleontological evidence provides some indication of the age of the esociform lineage and the history of its morphological diversity. If the fragmentary fossils described by Wilson et al. (1992) are correctly assigned to the esociformes, they show that this lineage originated as early as the late Cretaceous. More definite evidence of esociformes comes from nearly complete and well preserved fossil skeletons from the Paleocene of Alberta and Saskatchewan, which can be confidently assigned to *Esox* (Wilson, 1980) and from material of similar quality from the Eocene of Wyoming that may represent species of the subgenus *Kenoza* of *Esox* (Grande, 1999). This fossil evidence would imply that all four lineages leading to extant esociform genera had originated before the Paleocene if López et al.'s (2000) hypothesis of esociform intra-relationships is correct.

Some early esociform fossils exhibit characteristics that have been interpreted as in transition from the 'umbrid' to the esocid type (Sytchevskaya, 1976). However, Grande's (1999) summary of fossils of *Esox* shows that the paleontological evidence is insufficient to confidently infer the characteristics of the esociform ancestor that could be informative in determining the esociform sister group and inferring relationships among esociform lineages.

As part of ongoing efforts to better understand the phylogeny of actinopterygians and the evolution of esociformes, we have assembled a new data set consisting of DNA sequences from both the nuclear and mitochondrial genomes from most of the extant species of Esociformes and a wide sample of euteleosts to test previously proposed hypotheses of esociform intra- and interordinal relationships.

## Materials and Methods

*DNA sequences*

We obtained DNA sequences from representatives of the four extant esociform genera and a wide sample of possible outgroups. The taxonomic sample included at least one and more often several members of all the lineages that have been proposed as possible sister groups to the esociforms (e.g. 4 salmonoids, 8 neoteleosts, 13 protacanthopterygians; see Table 3.1 for the list of species and Genbank accession numbers of the sequences examined). This taxonomic sample is sufficient to test existing hypotheses of esociform relationships.

We sequenced a fragment from the 3' end of the single-copy *recombination activation gene -1* (RAG1) from the nuclear genome and two fragments from the 3' halves of the 12S and 16S ribosomal RNA genes from the mitochondrial genome. We used the PCR, gel electrophoresis, PCR-product purification, dye-deoxy chain termination, and automated sequence determination protocols to obtain the DNA sequences. For 12S, we used primers L1091 and H1478 (Kocher et al. 1989) and for 16S, 16Sar-L and 16Sbr-H (Palumbi et al. 1991). These primers amplify fragments of the 12S and 16S mitochondrial rRNA genes corresponding to positions 1508-1896 and 3009-3588 in the *Oncorhynchus mykiss* mitochondrial genome, respectively (Genbank accession NC001717; Zardoya et al., 1995). For RAG1, we used primers designed to amplify the more slowly evolving 3' half of the

gene (approximately 1400bp). The primer names and sequences are: RAG1F1- CTG AGC TGC AGT CAG TAC CAT AAG ATG T; RAG1R1- CTG AGT CCT TGT GAG CTT CCA TRA AYT T; RAG1R2- TGA GCC TCC ATG AAC TTC TGA AGR TAY TT; and RAG1R3- GTC TTG TGS AGG TAG TTG GT. The primer set RAG1F1 – RAG1R1 target the region spanning between nucleotide positions 2215 and 3772 of the Onchorhynchus mykiss RAG1 sequence accessioned in Genbank (U15663). The primer RAG1R2 is an alternative reverse primer that we used when RAG1R1 produced unsatisfactory results and the primer RAG1R3 is an internal primer that we used to obtain the full DNA sequence of the fragment amplified by the external primers.

We aligned the RAG1 sequences manually using the amino acid translation to guide the placement of the 10 amino acid insertions/deletions and the two introns (only found in *Argentina* and *Bathylagus*). We used the program Clustal X (Thompson et al., 1997) to align the 12S and 16S sequences and manually edited the resulting alignments guided by the secondary structure of these molecules as inferred by Wang and Lee (2002) and Waters et al. (2000), respectively. To edit the alignment in regions of high variability we used conserved stem regions as anchoring points. We discarded regions where the amount of length variation was very high and the resulting alignment would likely contain invalid assertions of homology.

We compared the base composition of the sequences using the chi-square test implemented in PAUP* version 4.0b10 (D. Swofford, unpubl) to determine the potential for artifacts in phylogenetic reconstruction that may result from convergence in base composition bias between taxa.

To obtain optimal maximum parsimony (MP) and minimum evolution (ME) trees from each data set we conducted 1000 replicate heuristic searches with random taxon

addition and random starting trees with tree-bisection-reconnection (TBR) branch swapping and saving the optimal tree from each replicate. We determined the bootstrap indices of support based on MP and ME criteria. In these bootstrap analyses, we conducted a full heuristic search for the optimal tree of 1000 bootstrap pseudoreplicates. Each heuristic search invoked TBR branch swapping. In MP bootstrap, each heuristic search consisted of two replicate searches with random taxon addition sequence and random starting tree. We present the 50% majority rule consensus tree of optimal trees from these searches of bootsrap pseudoreplicates. We also conducted a heuristic search for the optimal maximum likelihood (ML) tree of each data set to determine if this analysis produced conflicting results. To select the substitution model for ME and ML we used the routine devised by Posada and Crandall (1998) as implemented in Modeltest v. 3.06 (D. Posada, unpubl). We used the program PAUP* v. 4.0b10 (D. Swofford, unpubl) to conduct all other analyses.

We compared the phylogenetic hypothesis supported by the sequences with those that have been previously proposed using both parsimony and likelihood based tests (Kishino and Hasegawa, 1989; Shimodaira and Hasegawa, 1999) implemented in PAUP* v.4.0b10. In addition, we visually inspected the aligned RAG1 amino acid translation in search of substitutions that could be interpreted as synapomorphies of the clades supported by alternative hypotheses of esociform phylogeny.

*Review of evidence supporting existing hypothesis*

We critically examined the data that support previously published hypotheses of esociform relationships. We focused this evaluation on the supporting evidence offered by Nelson (1972), Wilson and Veilleux (1982: summarized in pp. 350-351) and Johnson and Patterson (1996: Sections II and X and, Appendix 4) because these studies offer explicit cladistic interpretations of the data. Our review of the evidence consisted of checking for

consistent and accurate character state coding and examination of cleared stained specimens (see Materials Examined) when needed to aid these assessments. We then evaluated whether or not issues detected in our review of the data could have a detrimental effect on the support given to phylogenetic hypothesis for esociforms.

## Results

### *RAG1 sequences*

The alignment of the RAG1 sequences used in this study consists of a fragment coding for 475 amino acid sites from the carboxyl half of the molecule (*Oncorhynchus mykiss*, U15663; Hansen and Kaattari, 1995). This fragment is located downstream of the intron RAG1b according to the nomenclature of Venkatesh et al. (1999). The DNA sequence alignment includes 1351 nucleotide sites after we eliminated sites with missing or ambiguous data. The sites eliminated include gaps created in the alignment by insertions/deletions in the amino acid sequence (10 residues) and two previously undescribed RAG1 introns that we discovered in the sequences of *Argentina sialis* and *Bathylagus ochotensis*. One of the two introns discovered in this study is shared by *B. ochotensis* and *A. sialis*, the only two argentiniforms in our taxonomic sample. The other intron was only found in *B. ochotensis* among the taxa sampled in this study. The distribution of amino acid insetions/deletions in our taxonomic sample was not phylogenetically informative so we discarded the sites involved.

When all sites in the RAG1 sequence are considered, the null hypothesis of base composition stationarity among the species in this study is rejected using the chi-square test implemented in PAUP* ($p<<0.0001$). Examination of base composition of the RAG1 sequences by codon position categories shows that the source of the heterogeneity is at the

third codon position sites. The null hypothesis of base composition stationarity at third codon position sites is strongly rejected ($p<<0.0001$). Among the taxa represented in this study, the combined proportion of G's and C's (GC content), varies widely and continuously in a range from 0.46 to 0.95. *Ophichthus gomesii*, *Gonostoma bathyphilum* and the two galaxiids in our sample (*Brachygalaxias bullocki* and *Galaxias fasciatus*) have GC content higher than 0.90 at third codon position sites. The ranges of GC content of osmeriforms (0.71-0.94), salmonoids (0.77-0.79) overlap and that of esociformes (0.65-0.74) overlaps with osmeriformes but not with salmonoids.

The chi-square test fails to reject the null for first and second codon position sites ($p=1.0$). At these sites, the magnitude of the range of GC content spanned by the taxa in this study is much reduced in comparison to that observed at third codon positions (0.11, 0.06, and 0.49 at $1^{st}$, $2^{nd}$ and $3^{rd}$ codon position sites, respectively). Because salmonoids and esociformes do not show shared distinct patterns of base composition, it is unlikely that the sister group relationship between salmonoids and esociforms supported by these sequences is an artifact of convergent base composition.

A plot of observed uncorrected transitions against transversions shows a deviation from a homogeneous relationship between transitions and transversions for RAG1 sequences (Fig 3.2). This deviation may be the result of changes in the substitution process and/or substitution saturation. The points in the plot become greatly dispersed at values above 6% transversions and 10% transitions. The extensive dispersion of points indicates heterogeneity in the substitution process among the taxa represented in the study. This observation is in agreement with the implications of the base composition heterogeneity and variability in GC content at third codon position sites described above.

There are six shortest trees under parsimony (5266 steps, CI 0.283, RI 0.532). The strict consensus of these trees is in general agreement with the MP bootstrap majority rule consensus tree. Both, MP and ME majority rule bootstrap trees (Fig 3.3A and B, respectively) strongly support the monophyly of the esociformes, the sister group relationship of salmonoids and esociforms and the esociform intra-ordinal relationships proposed by López et al. (2000). Also, the taxa in this study representing the argentinoids (2 species), eurypterigians (8 species), clupeiformes (2 species), galaxiids (2 species), ostariophysans (7 species), osmeroids (5 species), retropinnids (2 species), salmonoids (4 species) and, stomiiforms (2 species) form well supported clades. Support for the Euteleostei (without Ostariophysi) also is strong (70-93%); but relationships among euteleost outgroup taxa are not well-resolved, presumably due to poor taxonomic sampling. For example, the order Osteoglossiformes, represented by *Hiodon alosoides* and *Osteoglossum bicirrhosum*, receives weak to no support. The elopomorphs *Albula vulpes*, *Megalops atlanticus* and, *Ophichthus gomesii* variously form a weakly supported clade or are found to be paraphyletic. The base substitution model that best fits the data is the general time reversible (GTR) with among site rate variation (ASRV) and invariant sites (I) with the following parameter values: base frequencies (A 0.2253, C 0.2886, G 0.2848, T 0.2014), substitution rates ($r_{AC}$ 1.4203, $r_{AG}$ 3.4727, $r_{AT}$ 1.6624, $r_{CG}$ 0.8849, $r_{CT}$ 4.7203), gamma distribution shape parameter ($\alpha$) 1.1595, and proportion of invariant sites ($p_{inv}$) 0.384.

Parsimony and likelihood tree comparison tests indicate that alternative hypotheses of esociform and basal euteleost phylogeny represent significantly poorer tree topologies for the RAG1 sequences (Table 3.2). A visual inspection of the amino acid translation of these sequences revealed residues that may represent synapomorphies of: (1) the salmonoid + esociform clade (valine to threonine at 777, proline to cysteine or arginine at 906, and

threonine to valine at 958 of the *Oncorhynchus mykiss* translated sequence with Genbank accession no. AAA80281), (2) the clade composed of *Dallia* + *Esox* + *Novumbra* (valine to isoleucine at 850, glutamine to asparagine or aspartic acid at 852, AAN codon to glutamine at 874, and glutamine to lysine at 923), and (3) the clade composed of *Esox* + *Novumbra* (threonine to asparagine 641). We could not detect any amino acid substitutions that would be best interpreted as shared derived residues of an osmeriform+salmonoid clade or the Umbridae or Umbrinae.

*Mitochondrial ribosomal DNA sequences*

The alignment of the 12S sequences includes 302-312 nucleotides corresponding to sites 1551 through 1856 on the *Oncorhynchus mykiss* mitochondrial genome (NC001717; Zardoya et al., 1995). This region includes stems 27 through 40 of the secondary structure model presented by Wang and Lee (2002). Seven of the loops included in the sequenced region show extreme variability; thus we considered the homology of the sites implicit in the alignment suspect and report results of analyses with and without these hypervariable regions.

The alignment of 16S sequences includes 358-399 nucleotides corresponding to sites 3102 through 3479 of the *O. mykiss* mitochondrial genome (NC001717; Zardoya et al., 1995). This region is a portion of the 3' half of the 16S rDNA gene and includes the sites between loops E27 and G16 as presented in Waters et al. (2000). Three loops included in the sequenced region show extreme variability so we gave them the same treatment described above for the 12S hypervariable regions. For the 12S, 16S and combined mitochondrial ribosomal DNA (mt rDNA) sets of sequences, the chi-square test of base composition fails to reject the null hypothesis of stationarity.

Because the rDNA sequences are relatively short and given that they share many biological characteristics (e.g. mitochondrial inheritance, transcribed but not translated components of the ribosome), we combined these two genes in all phylogenetic analyses. There are 65 shortest trees for the mt rDNA sequences (1290 setps, CI 0.325, RI 0.516). The MP bootstrap majority rule consensus trees of the mt rDNA sequences with and without considering the hypervariable regions are largely unresolved. Clades such as esociformes, salmonoids and osmeroids receive moderate to strong support as indicated by bootstrap values. The osmeroid clade supported (98%) by parsimony includes the osmerids, *Plecoglossus*, *Salangichthys* and the retropinnids. The ostariophysans are not resolved as a monophyletic group and all higher relationships, with the exception of the Euteleostei without Ostariophysi (50 – 67%) remain unresolved.

The tree based on ME analysis using corrected distances supports similar results (Fig 3.4). There is little resolution of higher-level relationships, ostariophysan monophyly is not supported and strong bootstrap support is largely restricted to trivial groupings with the exception of the osmeroid clade described above, which is strongly supported (97%). When hypervariable sites are considered esociforms and salmonoids form a clade with weak support (63%), but when these sites are removed, the support for this group is weaker (56%). The Euteleostei without Ostariophysi is weakly supported (67%) when all sites, including those from hypervariable regions are considered. The base substitution model that best fits the data is the GTR + ASRV + I with the following parameter values: base frequencies (A 0.3186, C 0.2407, G 0.2038, T 0.2369), substitution rates ($r_{AC}$ 3.1199, $r_{AG}$ 10.2036, $r_{AT}$ 2.6920, $r_{CG}$ 1.0442, $r_{CT}$ 18.6542), $\alpha$ 0.6116, and $p_{inv}$ 0.3377.

Under ML, esociforms and salmonoids are sister groups both when all mt rDNA sites are considered and when the sites in the hypervariable regions are removed. Another

grouping found in both of these ML trees is the one containing all the eurypterygian taxa included in this study. Generally, the inferred length of terminal branches relative to internal ones is large but, not surprisingly, short terminal branches are inferred among osmeroids, salmonoids and esociformes; the groups that were most densely sampled in this study. Generally, tree topology comparison tests suggest that alternative hypothesis of esociform and lower euteleost phylogeny are inconsistent with these sequences. However, the Shimodaira-Hasegawa test fails to reject a sister group relationship between osmeroids and salmonoids (Table 3.2).

*Combined DNA sequences*

The null hypothesis of base composition stationarity is rejected for the combined sequences ($p<<0.0001$), which indicates that the deviation from base composition homogeneity at third codon position sites of RAG1 is extensive.

There is one MP parsimony tree of the combined RAG1 and mt rDNA data set (6606 steps, CI 0.289, RI 0.524). The bootstrap majority rule consensus tree from the combined sequences excluding hypervariable regions show very strong support for the following groups (Fig 3.5A): Euteleostei minus Ostariophysi (91%), Ostariophysi (99%), Osmeridae+Salangidae+Plecoglossidae+Retropinnidae (100%), Eurypterygii (100%) and Esociformes+Salmonoids (100%). Within Esociformes, the results of this analysis are in agreement with the inter-generic hypothesis of López et al. (2000). All but one of the esociform clades of that hypothesis appear in more than 99% of the bootstrap pseudoreplicates. The exception concerns the tree branch that defines a monophyletic *Esox* to the exclusion of *Novumbra*. This branch appears in only 69-80% of the trees from bootstrap pseudoreplicates. Within *Esox*, this analysis supports the monophyly of the subgenera *Esox* and *Kenoza* as defined by Nelson, 1972.

The results of ME analyses of the combined data set are in general agreement with those based on parsimony (Fig 3.5B). The clades described above also are strongly supported by analyses based on corrected genetic distances. In addition, a clade composed of argentinoids, osmeroids and stomiiforms is strongly supported (91%) and a clade composed of osmeroids and stomiiforms receives marginal support (63%). The base substitution model that best fits the data is the GTR + ASRV + I with the following parameter values: base frequencies (A 0.2503, C 0.285, G 0.2683, T 0.1964), substitution rates ($r_{AC}$ 1.3218, $r_{AG}$ 3.5285, $r_{AT}$ 1.622, $r_{CG}$ 0.8839, $r_{CT}$ 5.528), $\alpha$ 0.9834, and $p_{inv}$ 0.381.

The ML tree based on the combined data set contains the salmoniform+esociform, the osmerid+salangid+plecoglossid+retropinnid and the euteleost minus ostariophysan clades found in other analysis described above. It differs from parsimony and distance analyses of the same data in the placement of argentinoids and galaxiids as sequential sister groups of eurypterygians. The results of tree comparison tests based on the combined data set mirror those based on RAG1 sequences alone.

*Evidence supporting the sister group relationship of Neoteleostei and Esociformes*

The sister-group relationship of esociforms and neoteleosts is based on four characters, none of which represents an uncontradicted synapomorphy. This relationship is recognized as tentative (Johnson and Patterson, 1996). The following observations apply to those four characters: (1) The presence of Fink's (1981) Type 4 tooth attachment in some esociformes, some stomiatiforms and eurypterygians needs further examination because there are characteristics of the condition in *Esox* that may indicate a different origin for the depressible teeth in esociformes (Fink, 1981; p. 183). (2) The loss of Uroneural 3 occurs among membes of other basal euteleost lineages such as alepocephaloids, galaxiids and retropinnids. (3) Among neoteleosts, the presence of scales on the cheek and operculum, is

restricted to the eurypterygians thus in Johnson and Patterson's (1996) analysis this trait is reconstructed as a synapomorphy of the esocoid+neotelost clade only when forward changes are favored (p. 313: Fig 23B). It is noteworthy that under the coding of this trait given in their Appendix 4 (neoteleosts are coded as ambiguous) the trait cannot be considered synapomorphic. Finally, (4) acellular bone also occurs in osmeroids.

*Evidence supporting the sister group relationship of Salmonoidei and Osmeroidei*

We reviewed the evidence supporting the Salmoniformes (Salmonoidei + Osmeroidei) of Johnson and Patterson (1996) because our molecular data yield strong support for a clade formed by Esociformes and the salmonoids to the exclusion of the osmeroids. Johnson and Patterson (1996) found 11 characters that support their Salmoniformes. Of these 11 potential synapomorphies, four (nos. 4, 7, 18 and 37 of Appendix 4; Johnson and Patterson, 1996) are present in the putatively apomorphic state in at least one esociform. The fusion of dermethmoids and supraethmoids (no. 1) is difficult to code in esocoids because the homologies of the components of their ethmoid skeleton (proethmoids with an anterior endoskeletal ossification) are unclear. Johnson and Patterson (1996) considered the presence of a dermal component with an anterior endoskeletal thickening in the proethmoids of some esociforms to represent the 'fused' condition of the ethmoid skeleton; however we do not find justification to equate these paired anterior endoskeletal ossifications with the supraethmoid of other euteleosts. In the absence of clear homologues to the supraethmoids in the ethmoid skeleton of esociformes it is impossible to determine the appropriate state of this character in the group. For another character (no. 16), the retention of the upper pharyngeal toothplate 5, the condition ascribed to esociformes was determined by observations of specimens of *Esox americanus*, which is a member of the most derived esociform genus according to the molecular evidence. Because *Esox* may be a

derived esociform and because the gill arches of all species of *Esox* must accommodate the elongated snout that is autapomorphic for the genus, the anatomy of the association between tooth-bearing plates and gill arch elements in *Esox* may not be representative of the ancestral esociform condition. Our own observations of upper pharyngeal toothplates in salmonids, osmerids and 'umbrids' revealed similar morphologies.

The ancestral character state of the ossification of epipleurals (no. 21) in osmeroids is problematic (Johnson and Patterson, 1996; p. 279). Uroneural 2 is absent in all esocoids, so its placement relative to uroneural 1 cannot be coded (no. 29). The value of the presence of nuptial tubercles and diadromy as salmoniform synapomorphies is debatable because both traits show homoplasy outside basal euteleosts (nos. 40 and 42, respectively).

Our review indicates that the evidence reported does not provide strong support for the exclusion of esociformes from a salmonoid-osmeroid clade. Of the 11 characters considered synapomorphic for Johnson and Patterson's Salmoniformes, we find that only the presence of rostro-caudal expansions of the last few neural and haemal spines (no. 24) is not problematic in this context.

*Evidence supporting the currently accepted hypothesis of relationships among esociforms*

Nelson's (1972; Fig.1B) hypothesis of esociform relationships receives support from characters proposed by Nelson (1972) and Wilson and Veilleux (1982). We reviewed this evidence because it contradicts the molecular-based hypothesis proposed by López et al. (2000) and supported in this study. In particular, the molecular evidence is inconsistent with a monophyletic Umbridae and Umbrinae (*Umbra+Dallia*).

According to Nelson (1972; p. 21), the monophyly of Umbridae is supported by four characters that are all related to the loss or reduction of components of the cephalic sensory system. The monophyly of Umbrinae is supported also by four characters that describe loss

or reduction of components of the cephalic sensory system. Two of these, presence or absence of the posterior part of the infraorbital canal and the number of pores on the preopercular canal, are secondary character state transformations of traits offered in support of umbrid monophyly. We did not find anything problematic in the coding of these characters but it is important to note that: (1) they all describe similar transformations (i.e. reduction or loss) of different components of a single system, thus they may not satisfy the criterion of independence and (2) the ordered character transformation implied by the reductive trend asserts a directional model of evolution that needs justification before it can be used to support a 'phylogenetic sequence'.

Wilson and Veilleux (1982) assumed the monophyly of Umbridae in their study of umbrid relationships. They listed four characters that distinguished umbrids from esocids (p. 346-347) but did not clarify whether or not these represented umbrid synapomorphies. The plesiomorphic state for esociformes for two of these characters (number of branchiostegals and number of vertebrae) is difficult to determine with any confidence. One of the remaining two characters (the number of infraorbitals) refers to the character complex described by Nelson (1972). The fourth umbrid trait listed by Wilson and Veilleux (1972) is the absence of supratemporals and intercalars. In their Appendix, Wilson and Veilleux (1982; Set I, p. 350) list 13 "shared derived characters" that support the monophyly of Umbrinae. Two of these (nos. 3 and 12) are miscoded according to the information provided in the text and illustrations and also according to our own observations of cleared and stained specimens. The coding of the presence of a knob on the proethmoid (no. 1 from Set I, p. 350) is problematic for two reasons: the plesiomorphic condition cannot be determined because the shape of the proethmoids of all esociformes is different from that observed in other basal euteleosts with proethmoids and the description of the character (p. 326) does not specify to

what extent the condition described for *Umbra* and *Dallia* (the putative apomorphy) differs from that reported for *Esox*, which also shows an anterior thickening of the proethmoid. Two other putative synapomorphies (no. 4 and 6) are coded as reduced in *Dallia* and *Umbra* but the structures to which those characters refer are absent in *Umbra* (no. 4) and *Dallia* (no. 6), thus raising the question of whether or not a missing structure may be equated to a reduced structure in the context of character state coding; further we did not observe markedly reduced ectopterygoids (no. 6) in *U. limi* and *U. pygmae*.

Four characters (nos. 8-11) refer to different aspects of the anatomy of the caudal skeleton. Three of them (nos. 9-11) show within genus variation and are inconsistently coded when compared to illustrations presented by Rosen (1974) and Wilson and Veilleux (1982), and to our own observations. Of these four characters of the caudal skeleton, the reduced difference in size between hypural 1 and hypurals 2 and 3 in *Dallia* and *Umbra* compared to *Esox* and *Novumbra* (no. 8) is the only one consistent with the illustrations and our observations.

Finally, the number of radials in the pectoral girdle (no. 13) of *Umbra* is given as less than four, but we have observed four in all the specimens we checked. Some specimens have the two ventral most radials fused at their ends, but in all cases the presence of four distinct elements is clear. Even if we ignore this observation, we consider it a strained argument to propose the highly modified unossified radial plate of *Dallia* and its rather ordinary homologue in *Umbra* as evidence of a shared character state transformation.

Our review of Wilson and Veilleux's (1982) evidence revealed that only four of the 13 putative synapomorphies of the Umbrinae are not problematic. Two of these (nos. 2 and 5) are taken from Nelson (1972) and were discussed above. The other two characters are (no.

7) the loss of the basihyal toothplate, a reductive trait, and (no. 8) a reduced difference in size between hypural 1 and hypurals 2 and 3.

In summary, we did not find a strong body of evidence supporting the above reviewed hypotheses of esociform relationships. Our goal here was not to reanalyze the morphological data based on a reinterpretation of the evidence, but to determine the degree of confidence that phylogenies base on those data should receive. Based on this review of the published evidence, we did not find a compelling case for favoring any of these hypotheses over new proposals.

## Discussion

### *Esociforms among basal euteleosts*

Our analyses of the new molecular data show strong support for a sister group relationship between salmonoids and Esociformes. The esociform – salmonoid sister group relationship has been previously proposed by Fink (1984) and Williams (1987), who described characters of cheek musculature in support of this proposal but deemed the evidence too weak to be conclusive. Previous molecular studies offer corroborative evidence for the esociform+salmonoid clade. The nuclear rDNA sequences reported by Lê et al. (1989) and the growth hormone sequences analyzed by Bernardi et al. (1993) offer support for this group. Zaragüeta-Bagils et al. (2002) summarized evidence from six data sets of DNA sequences from teleosts to uncover phylogenetic relationships that could be accepted with confidence based on agreement between different sources of evidence. Although none of the data sets they examined includes a significant representation of protacanthopterygian taxa, in all cases where both esociforms and salmonoids were represented, these two lineages were placed as sister groups with relatively strong measures of support. A more recent study

examined the validity of the Protacanthopterygii concept using whole mitochondrial genome DNA sequences (Ishiguro et al., in press) from a wide sample of teleost taxa. The two esociforms (*Esox lucius* and *Dallia pectoralis*) included in that study were supported as the sister group of the salmonoids. We consider the general consensus of the molecular evidence as a strong indication that the grouping of salmonoids and esociformes in a monophyletic clade accurately represents the history of these two lineages.

*Esociform intrarelationships*

Results from all of our analyses are in complete agreement and offer strong support for the hypothesis of esociform inter-generic relationships proposed by López et al. (2000). The bootstrap analyses in general show well-supported clades among the esociformes. The only inconsistently supported node grouping esociform taxa is the one that defines the monophyly of *Esox* to the exclusion of *Novumbra* (e.g. Fig 3.4).

The genetic distances among the most divergent species of *Esox* and between those species and *Novumbra hubbsi* are of similar magnitudes. This observation may be explained by: (a) substitution saturation in the sequences we examined, (b) a markedly different rate of molecular evolution in species of *Esox* compared to that of *Novumbra*, or (c) a similar time of origin for the lineages of *Novumbra* and the subgenera *Esox* and *Kenoza*. Substitution saturation is an unlikely explanation because the levels of divergence among these taxa are much lower than those observed in other comparisons and they are found well within range of linear change in the substitution plot (Fig 3.2). Relative rate tests with *Dallia* as the reference outgroup point to a faster rate of substitution in the *Novumbra* lineage relative to lineages of *Esox*. While we cannot rule out a similar time of origin for all these lineages, differences in the rate of evolution evident in the sequences may be responsible for the inconsistent support for *Esox* monophyly. Within the Esocidae, our data strongly support the

division of the species of *Esox* into the subgenera *Kenoza* for the pickerels and *Esox* for the pikes and the muskellunge, and within the subgenus *Esox*, *E. lucius* and *E. reichertii* are supported as sister species as proposed by Nelson, 1972.

In conclusion, the molecular data presented here offer unambiguous support for the phylogenetic hypothesis proposed by López et al. (2000). Because our study includes a thorough sample of extant esociform diversity, a broad sample of potential sister lineages, and sequence data from different genes and genomes (i.e. nuclear and mitochondrial), a revised classification of these taxa seems warranted. The classification proposed by López et al. (2000) may be accepted, however, the ranks given to the different groups in that classification may need revision once consensus on the placement of the salmonoid + esociform clade develops.

*Previously published hypotheses of esociform phylogeny*

Our review of the morphological evidence used to erect previously published hypotheses of esociform inter- and intra-ordinal relationships shows that the support for these phylogenetic inferences is weak. The placement of esociforms as the sister group of the neoteleosts is based on four characters that are difficult to accept as uncontested apomorphies shared by the ancestors of these two lineages. The alternative placement of esociforms as a basal lineage of euteleosts is based on the absence of evidence to place them elsewhere.

We also found problems with the evidence offered in support of Nelson's (1972) hypothesis of esociform relationships. This hypothesis rests largely on a putative reductive trend in different aspects of the cephalic sensory system. A total of four character state transformations involving losses of different components of the cephalic sensory system were used by Nelson to derive his esociform hypothesis.

In a recent examination of the evidence supporting hypotheses of Cetacean affinities, Naylor and Adams (2001) demonstrate the potential perils of extracting multiple characters from various aspects of a given anatomical component. Coding of the reductive trend in the cephalic sensory system of esociforms as multiple character transformations clearly fits the criteria outlined by Naylor and Adams (2001) for potentially problematic usage of evidence. While Nelson's (1972) esociform phylogeny hypothesis originally rested solely on the characters of cephalic sensory system when it was first proposed, it appeared to find corroboration in Wilson and Veilleux's (1982) findings. But a critical problem with their work is that they failed to consider the possible paraphyly of the 'Umbridae.' Therefore, Wilson and Veilleux's (1982) treatment of the data was biased by their explicit assumption that *Esox* represents the basal esociform genus (in other words, this critical premise was not demonstrated by evidence but asserted *a priori*). In addition, many of the characters (9 of 13) supporting their findings are incorrectly coded or the rationale for the reported coding is not satisfactorily justified. As a result, we consider existing hypothesis of esociform relationships tentative and present new data to evaluate alternative hypothesis.

*Other phylogenetic considerations*

Although our taxonomic sampling in this study was designed to address esociform relationships among protacanthopterygians, some of the results concerning other aspects of the phylogeny of basal euteleosts are worth highlighting.

McDowall, tentatively, (1969) and later Rosen (1974) suggested a close relationship between Southern (retropinnids) and Northern (osmerids) smelts. All of our analyses agree in placing the retropinnids as the sister group of the osmerid + salangid + plecoglossid clade to the exclusion of the two galaxiid species in our sample (Figs. 3,4, and 5). Clearly, our galaxiid sample is deficient and any hypothesis derived from these results must be further

tested. However, the same hypothesis was obtained by a detailed analysis of mtDNA data based on a taxonomic sample better suited to address retropinnid affinities (Waters et al., 2002).

One potential problem with the support given to the placement retropinnids by our sequence data is that the two galaxiids and the two retropinnids in our study show sharply contrasting patterns of base composition. The differences are more marked in the RAG1 sequences but they are also evident to a lesser extent in the mt rDNA data. In all cases, the sequence base composition of the two retropinnids is more similar to that of the five osmeroids than to that of the two galaxiids, therefore it is possible that the retropinnid + osmeroid clade is an artifact of base composition similarity. Alternatively, similar base composition may be a shared derived trait and as such indicative of phylogenetic affinity.

Within the osmeroid clade, our results do not support the placement of the salangids among osmerids as proposed by Johnson and Patterson (1996). Again, our taxonomic sample is inappropriate to produce a strong inference on this matter, but we consistently find very strong support for the monophyly of osmerids. Saruwatari et al. (2000) also reported strong support for monophyletic Osmeridae and Salangidae from analyses based on 16S sequences from a taxonomic sample that included five osmerid and three salangid genera, as well as *Plecoglossus altivelis*. The comparison tests we conducted showed that Saruwatari et al.'s (2000) hypothesis of osmeroid relationships was consistent with our data, while that of Johnson and Patterson (1996) was not (Table 3.2). On the other hand, the four uncontradicted morphological synapomorphies placing salangids and the osmerid genus *Mallotus* as sister groups in Johnson and Patterson's (1996: Fig. 19 and Appendix 1) analyses must be reconciled with this result. As our current understanding of the evolution of

morphological traits is not sufficiently developed to confidently assert the relative value of putative synapomorphies, it seems that the evidence as a whole supports osmerid monophyly.

While our results concerning osmeroid, retropinnid and galaxiiid relationships are in agreement with other molecular studies relevant to these questions, a strong conclusion will only be possible from a thorough critical examination of the evidence supporting existing hypotheses and the production of data sets from a sample of taxa designed to permit strong conclusions (i.e. complete or near complete representation of ingroup diversity and broad and well selected outgroup representatives).

Finally, although the DNA sequences we obtained for this study were informative about esociform relationships and promise to be informative about other relationships involving similar levels of divergence, there are deeper aspects of the basal euteleostean phylogeny for which we could not discern any relevant relationships. A denser taxonomic sample may show whether this lack of resolution is the result of poor sampling or the absence of phylogenetic information in the sequences. We suspect that to gain a better understanding of some of the more problematic aspects of euteleost evolution, it will be necessary both to produce the appropriate taxonomic samples and, in some cases, to employ new sources of evidence such as the distribution of introns (Venkatesh et al., 1999), conserved insetion/deletions (Venkatesh et al., 2001) and mobile genetic elements (e.g. SINE's; Shedlock and Okada, 2000). For example, the two novel RAG1 introns we found in the two argentinoids in this study may prove to have a phylogenetically informative distribution among lineages of this group. However, a critical flaw we have observed in many of the phylogenetic studies employing novel approaches is that in the rush to produce results, the quality of the taxonomic sampling has suffered, which makes the reliability of the conclusions they support impossible to gauge (e.g. Venkatesh et al. 2001).

## Material Examined

The following cleared and stained specimens were used to check some of the morphological characters used to support existing hypotheses of esociform relationships (see Discussion; Institution abbreviations follow Leviton et al., 1985): Dallia pectoralis, Bethel, AK, UMMZ 164848; Anchorage, AK, unaccesioned; Esox americanus, Livingston Co., MI, UMMZ 202358; Wilson Co., NC, unaccessioned; E. lucius, Chippewa R., WI, FMNH 18090; Spirit Lake Hatchery, IA, unaccessioned; E. masquinongy, Spirit Lake Hatchery, IA, unaccessioned; Novumbra hubbsi, Grays Harbor, WA, UMMZ 179398, UMMZ 187427; Umbra krameri, Lake Pantelimon, Romania, UMMZ 185076; U. limi, Mackinac, MI, UMMZ 137450; Jackson Co., IA, unaccessioned; U. pygmaea, Nansemond & Norfolk, VI, UMMZ 164967; Wilson Co., NC, unaccessioned.

## Acknowledgements

We thank the NSF (grant DEB 9985045 to G.O.) and Iowa State University for financial support to conduct this research. JAL thanks G. Naylor for guidance and encouragement throughout this study. G. Naylor generously provided the RAG1 sequence information needed for primer design. We also thank the following colleagues for providing samples for this study: G. Bernardi, J. Billerbeck, R. Kopf, G. Lecointre, T. Pietsch, T. Saruwatari, L. Suneetha, E. Verheyen, and J. Waters.

## Literature Cited

BERNARDI, G., G. D'ONOFRIO, AND S. CACCIO. 1993. Molecular phylogeny of bony fishes, based on the amino acid sequence of the growth hormone. J. Mol. Evol. 37:644-649.

CROSSMAN, E. J., and P. RAB. 2001. Chromosomal NOR phenotype and C-banded karyotype of Olympic mudminnow, *Novumbra hubbsi* (Euteleostei: Umbridae). Copeia 2001:860-865.

FINK, W. L. 1981. Ontogeny and phylogeny of tooth attachment modes in actinopterygian fishes. J. Morph. 167:167-184.

—. 1984. Basal euteleosts: relationships, p. 202-206. *In:* Ontogeny and systematics of fishes. Vol. Spec. Publ. No. 1. H. G. Moser, W. J. Richards, D. M. Cohen, M. P. Fahay, A. W. Kendall, and S. L. Richardson (eds.). American Society of Ichthyologists and Herpetologists, Lawrence, KS.

FINK, W. L., AND S. H. WEITZMAN. 1982. Relationships of the stomiiform fishes (Teleostei), with a description of *Diplophos*. Bull. Mus. Comp. Zool. 150:31-93.

GRANDE, L. 1999. The first Esox (Esocidae: Teleostei) from the Eocene Green River Formation, and a brief review of esocid fishes. J. Vert. Paleontol. 19:271-292.

GREENWOOD, P. H., D. E. ROSEN, S. H. WEITZMAN, AND G. S. MEYERS. 1966. Phyletic studies of teleostean fishes, with a provisional classification of living forms. Bull. Am. Mus. Nat. Hist. 131:339-456.

HANSEN, J. D., AND S. L. KAATTARI. 1995. The recombination activation gene 1 (RAG1) of rainbow trout (Oncorhynchus mykiss): cloning, expression, and phylogenetic analysis. Immunogenetics. 42:188-95.

ISHIGURO, N.B., M. MIYA AND M. NISHIDA. In press. Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the "Protacanthopterygii". Mol. Phyl. Evol.

JOHNSON, G. D., AND C. PATTERSON. 1996. Relationships of lower euteleostean fishes, p. 251-332. *In:* Interrelationships of fishes. M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson (eds.). Academic Press, San Diego.

KISHINO, H., AND M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of the Hominoidea. J. Mol. Evol. 29:170-179.

KOCHER, T. D., W. K. THOMAS, A. MEYER, S. V. EDWARDS, S. PÄÄBO, F. X. VILLABLANCA AND A. C. WILSON. Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. USA 86:6196-6200.

LÊ, H. L., R. PERASSO AND R. BILLARD. 1989. Phylogénie moléculaire préliminaire des 'poissons' basée sur l'analyse de sequences d'ARN ribosomique 28S. Seances Acad. Sci., Paris Ser. 3. 309: 493-498.

LÓPEZ, J. A., P. BENTZEN, AND T. W. PIETSCH. 2000. Phylogenetic relationships of esocoid fishes (Teleostei) based on partial cytochrome b and 16S mitochondrial DNA sequences. Copeia. 2000:420-431.

MCDOWALL, R. M. 1969. Relationships of galaxioid fishes with a further discussion of salmoniform classification. Copeia. 1969:797-824.

NAYLOR, G. J., AND D. C. ADAMS. 2001. Are the fossil data really at odds with the molecular data? Morphological evidence for cetartiodactyla phylogeny reexamined. Syst. Biol. 50:444-53.

NELSON, G. J. 1972. Cephalic sensory canals, pitlines, and the classification of esocoid fishes, with notes on galaxiids and other teleosts. Am. Mus. Novit. 2492:1-49.

PALUMBI, S. R., A. P. MARTIN, S. ROMANO, W. O. MCMILLAN, L. STICE AND G. GRABOWSKI. 1991. The simple fool's guide to PCR. Spec. Publ. Dept. Zool. University of Hawaii, Honolulu.

PARENTI, L. R. 1986. The phylogenetic significance of bone types in euteleost fishes. Zool. J. Linn. Soc. 87:37-51.

POSADA, D. AND K.A. CRANDALL. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14(9): 817-818.

ROSEN, D. E. 1973. Interrelationships of higher euteleostean fishes, p. 397-513. *In:* Interrelationships of fishes. P. H. Greenwood, R. S. Miles, and C. Patterson (eds.). Academic Press, London.

—. 1974. Phylogeny and zoogeography of salmoniform fishes and relationships of *Lepidogalaxias salamandroides*. Bull. Am. Mus. Nat. His. 153:265-326.

SARUWATARI, T., I. OOHARA, J. W. ORR, R. M. MCDOWALL, AND T. KOBAYASHI. 2000. Phylogeny of lower euteleosts reconstructed from mtDNA analysis. DNA polymorphism. 8:96-101.

SHEDLOCK, A. M., AND N. OKADA. 2000. SINE insertions: powerful tools for molecular systematics. Bioessays 22:148-160.

SHIMODAIRA, H., AND M. HASEGAWA. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114-1116.

SYTCHEVSKAYA, E. K. 1976. The fossil esocoid fishes of the USSR and Mongolia. Trudy Paleontologicheskogo Instituta, Akademiya Nauk USSR. 156:1-116.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN, AND D. G. HIGGINS. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research. 25:4876-82.

VENKATESH, B., M. V. ERDMANN AND S. BRENNER. 2001. Molecular synapomorphies resolve evolutionary relationships of extan jawed vertebrates. Proc. Natl. Acad. Sci. USA 98:11382-11387.

VENKATESH, B., Y. NING AND S. BRENNER. 1999. Late changes in spliceosomal introns define clades in vertebrate evolution. Proc. Natl. Acad. Sci. USA 96:10267-10271.

WANG, H.Y., AND S.C. LEE. 2002. Secondary structure of mitochondrial 12S rRNA among fish and its phylogenetic applications. Mol. Biol. and Evol. 19:138-148.

WATERS, J. M., J. A. LÓPEZ AND G. P. WALLIS. 2000. Molecular phylogenetics and biogeography of galaxiid fishes (Osteichthyes: Galaxiidae): Dispersal, vicariance, and the position of *Lepidogalaxias salamandroides*. Syst. Biol. 49:777-795.

WATERS, J. M., T. SARUWATARI, T. KOBAYASHI, I. OOHARA, R. M. MCDOWALL AND G. P. WALLIS. 2002. Phylogenetic placement of retropinnid fishes: Data set incongruence can be reduced by using asymmetric character state transformation costs. Syst. Biol. 51:432-449.

WILLIAMS, R. R. G. 1997. The phylogenetic relationships of the salmoniform fishes based on the suspensorium and its muscles. Unpublished Ph. D. Dissertation, University of Alberta, Edmonton.

WILSON, M. V. H. 1980. Oldest known *Esox* (Pisces: Esocidae), part of a new Paleocene teleost fauna from western Canada. Can. J. Earth Sci. 17:307-312.

WILSON, M. V. H., D. B. BRINKMAN, and A. G. NEUMAN. 1992. Cretaceous Esocoidei (Teleostei): Early radiation of the pikes in North American fresh waters. J. Paleontol. 66:839-846.

WILSON, M. V. H., and P. VEILLEUX. 1982. Comparative osteology and relationships of the Umbridae (Pisces: Salmoniformes). Zool. J. Linn. Soc. 76:321-352.

ZARAGUETA-BAGILS, R., S. LAVOUE, A. TILLIER, C. BONILLO, and G. LECOINTRE. 2002. Assessment of otocephalan and protacanthopterygian concepts in the light of multiple phylogenies. C.R. Biologies. 325:1-17.

ZARDOYA, R., A. GARRIDO-PERTIERRA, and J. M. BAUTISTA. 1995. The complete nucleotide sequence of the mitochondrial DNA genome of the rainbow trout, Oncorhynchus mykiss. J. Mol. Evol. 41:942-51.

**Table 3.1** Genbank accession numbers of the DNA sequences examined in this study

| | Genbank accession no. | | |
|---|---|---|---|
| Taxa | 12S | 16S | RAG1 |
| *Scaphirhyncus albus* | AY430247 | AY430229 | AY430198 |
| *Amia calva* | AB042952 | AB042952 | AY430199 |
| Osteoglossomorpha | | | |
| *Hiodon alosoides* | AY430248 | AY430230 | AY430200 |
| *Osteoglossum bicirrhosum* | AB043025 | AB043025 | AY430201 |
| Elopomorpha | | | |
| *Albula vulpes* | X99180 | X99179 | AY430202 |
| *Megalops atlanticus* | X99178 | X99177 | AY430204 |
| *Ophichthus gomesii* | AY430249 | AY430231 | AY430203 |
| Clupeomorpha | | | |
| *Engraulis japonicus* | AB040676 | AB040676 | AY430205 |
| *Pellona flavipinnis* | AY430250 | AY430232 | AY430206 |
| Ostariophysi | | | |
| *Chanos chanos* | AY430251 | AY430233 | AY430207 |
| *Danio rerio* | AC024175 | AC024175 | U71093 |
| *Pimephales promelas* | AY430253 | AY430235 | AY430210 |
| *Ictalurus nebulosus* | AY430252 | AY430234 | AY430209 |
| *Corydoras sp.* | U15271 | U15247 | AY430208 |
| *Gnathocharax steindachneri* | U33589 | U33624 | AY430211 |
| *Catoprion mento* | AF283911 | AF283932 | AY430212 |

**Table 3.1** (continued)

| Taxa | 12S | 16S | RAG1 |
|---|---|---|---|
| Osmeroidei | | | |
| *Brachygalaxias bullocki* | AY430266 | AF112328 | AY430219 |
| *Galaxias fasciatus* | AY430265 | AF112333 | AY430218 |
| *Retropinna tasmanica* | AY430263 | AF112342 | AY430216 |
| *Stokellia anisodon* | AY430264 | AF454843 | AY430217 |
| *Salangichthys microdon* | AY430267 | AY443566 | AY380539 |
| *Plecoglossus altivelis* | AY430261 | AY443567 | AY380536 |
| *Thaleichthys pacificus* | AY430262 | AY443568 | AY380537 |
| *Spirinchus thaleichthys* | AY430259 | AY430239 | AY430215 |
| *Hypomesus olidus* | AY430260 | AY443569 | AY380538 |
| Argentinoidei | | | |
| *Argentina sialis* | AY430258 | AY430238 | AY430228 |
| *Bathylagus ochotensis* | AY430257 | AY443570 | AY443564-5 |
| Salmonoidei (Johnson and Patterson, 1996) | | | |
| *Prosopium williamsoni* | AY430254 | AY430236 | AY430213 |
| *Thymallus thymallus* | AY430255 | AY430237 | AY430214 |
| *Salvelinus malma* | AY430256 | AF060445 | AY380535 |
| *Oncorhynchus mykiss* | L29771 | L29771 | U15663 |
| Esociformes | | | |
| *Esox masquinongy* | AY430274 | AY443571 | AY380543 |
| *Esox reichertii* | AY430277 | AY443572 | AY380545 |
| *Esox lucius* | AY430273 | AF060446 | AY380542 |

**Table 3.1** (continued)

| Taxa | 12S | 16S | RAG1 |
|---|---|---|---|
| *Esox niger* | AY430276 | AY443573 | AY380544 |
| *Esox americanus* | AY430275 | AY443574 | AY380541 |
| *Novumbra hubbsi* | AY430272 | AF060447 | AY380546 |
| *Dallia pectoralis* | AY430271 | AF060448 | AY380540 |
| *Umbra krameri* | AY430269 | AF060444 | AY380547 |
| *Umbra limi* | AY430268 | AF060443 | AY380548 |
| *Umbra pygmaea* | AY430270 | AF060442 | AY380549 |
| Neoteleostei | | | |
| *Vinciguerria sp.* | AY438704 | AY443575 | AY442363 |
| *Gonostoma bathyphilum* | AY438705 | AY443576 | AY438703 |
| *Notoscopelus kroeyeri* | AY430279 | AJ277964 | AY430221 |
| *Chlorophthalmus sp.* | AY430278 | AY430241 | AY430220 |
| *Regalecus glesne* | AF049728 | AF049738 | AY430222 |
| *Sargocentron punctatissimum* | AY430280 | AY430242 | AY430223 |
| *Menidia menidia* | AY430281 | AY430243 | AY430225 |
| *Lepomis macrochirus* | AY430284 | AY430246 | AY430227 |
| *Etheostoma caeruleum* | AY430283 | AY430245 | AY430226 |
| Trinectes maculates | AY430282 | AY430244 | AY430224 |

**Table 3.2** P-values of parsimony and likelihood tree comparison tests.

| Hypothesis | RAG1 | | | | mt rDNA | | | | Combined | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Parsimony | | Likelihood | | Parsimony | | Likelihood | | Parsimony | | Likelihood | |
| | W.S.[1] | K-H[2] | S-H[3] | K-H[4] | W.S. | K-H | S-H | K-H | W.S. | K-H | S-H | K-H |
| Esociformes - 1972[5] | <0.001 | <0.001 | NA | <0.001 | 0.052 | 0.033 | NA | 0.007 | <0.001 | <0.001 | NA | <0.001 |
| Osmeroids + Salmonoids[6] | <0.001 | <0.001 | <0.001 | NA | 0.007 | 0.018 | 0.171 | NA | <0.001 | <0.001 | <0.001 | NA |
| Lower euteleosts - 1996[7] | <0.001 | <0.001 | <0.001 | NA | 0.053 | 0.028 | 0.006 | NA | <0.001 | <0.001 | <0.001 | NA |
| Osmeroids – 1996[8] | <0.001 | <0.001 | NA | 0.001 | 0.004 | 0.002 | NA | 0.007 | <0.001 | <0.001 | NA | <0.001 |

[1] Winning-sites test (non-parametric) as implemented in PAUP*v.4.0b10

[2] Parsimony analog of test devised by Kishino and Hasegawa (1989) as implemented in PAUP*v.4.0b10

[3] likelihood test for *a posteriori* comparisons devised by Shimodaira and Hasegawa (1999) as implemented in PAUP*v.4.0b10

[4] likelihood test for *a priori* comparisons devised by Kishino and Hasegawa (1989) as implemented in PAUP*v.4.0b10

[5] López et al.'s (2000) hypothesis of esociform relationships compared with Nelson's (1972) hypothesis

[6] Sister group relationship between esociforms and salmonoids compared with sister group relationship between osmeroids and salmonoids

[7] Optimal tree compared with Johnson and Patterson's (1996: Fig. 23) hypothesis of lower euteleostean relationships

[8] Saruwatari's (2000: Fig. 3) hypothesis of osmeroid relationships compared with Johnson and Patterson's (1996: Fig. 19) hypothesis

**Figure Legends**

**Figure 3.1** Hypothesis of esociform relationships: (A) Johnson and Patterson, 1996; (B) Nelson, 1972; (C) López et al., 2000.

**Figure 3.2** Plot of the observed proportion of transitions vs. transversions for all pairwise species comparisons of RAG1 sequences. Comparisons involving esociform, salmonoid and osmeroid taxa are highlighted (◆) among all other comparisons (□).

**Figure 3.3** Majority-rule consensus tree from 1000 bootstrap pseudoreplicates of the RAG1 sequence data using (A) maximum parsimony and (B) ME from corrected pairwise genetic distances with substitution model parameter settings obtained with Modeltest (GTR+ASRV+I, parameter values given in the text). Bootstrap values are given above the branches on the trees; asterisks indicate values ≥ 98%.

**Figure 3.4** ME phylogram based on combined 12S+16S rDNA sequence data using corrected pairwise genetic distances with substitution model parameter settings obtained with Modeltest (GTR + ASRV +I, parameter values given in the text). Bootstrap values are given for the branches that appear in the ME bootstrap consensus tree (1000 pseudoreplicates); asterisks indicate values ≥ 98%.

**Figure 3.5** Majority rule consensus tree from 1000 bootstrap pseudoreplicates of the combined RAG1 and mt rDNA sequence data using (A) maximum parsimony and (B) ME from corrected pairwise genetic distances with substitution model parameter settings obtained with Modeltest (GTR + ASRV +I, parameter values given in the text). Bootstrap values are given above the branches on the trees; asterisks indicate values ≥ 98%.

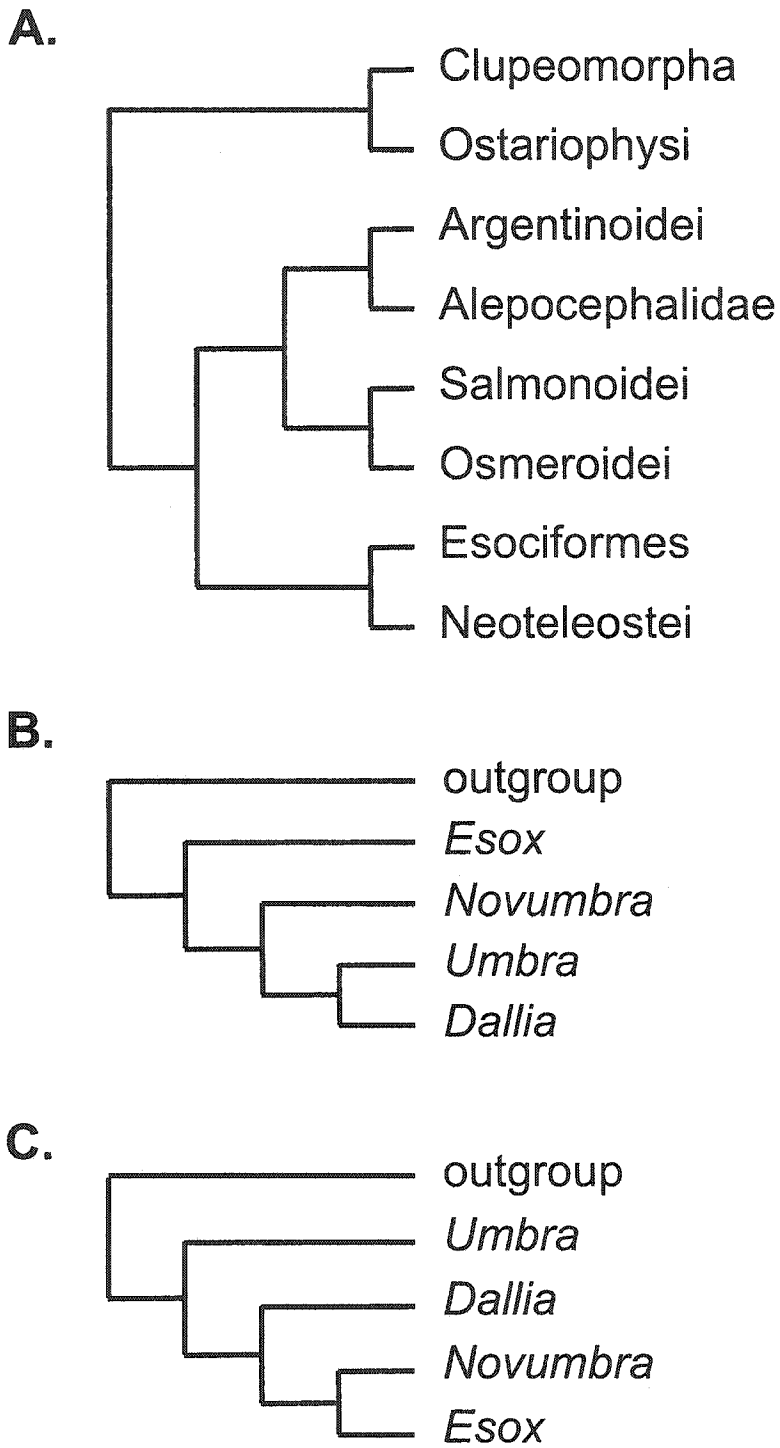
A.
Clupeomorpha
Ostariophysi
Argentinoidei
Alepocephalidae
Salmonoidei
Osmeroidei
Esociformes
Neoteleostei
B.
outgroup
Esox
Novumbra
Umbra
Dallia
C.
outgroup
Umbra
Dallia
Novumbra
Esox

**Figure 3.1**

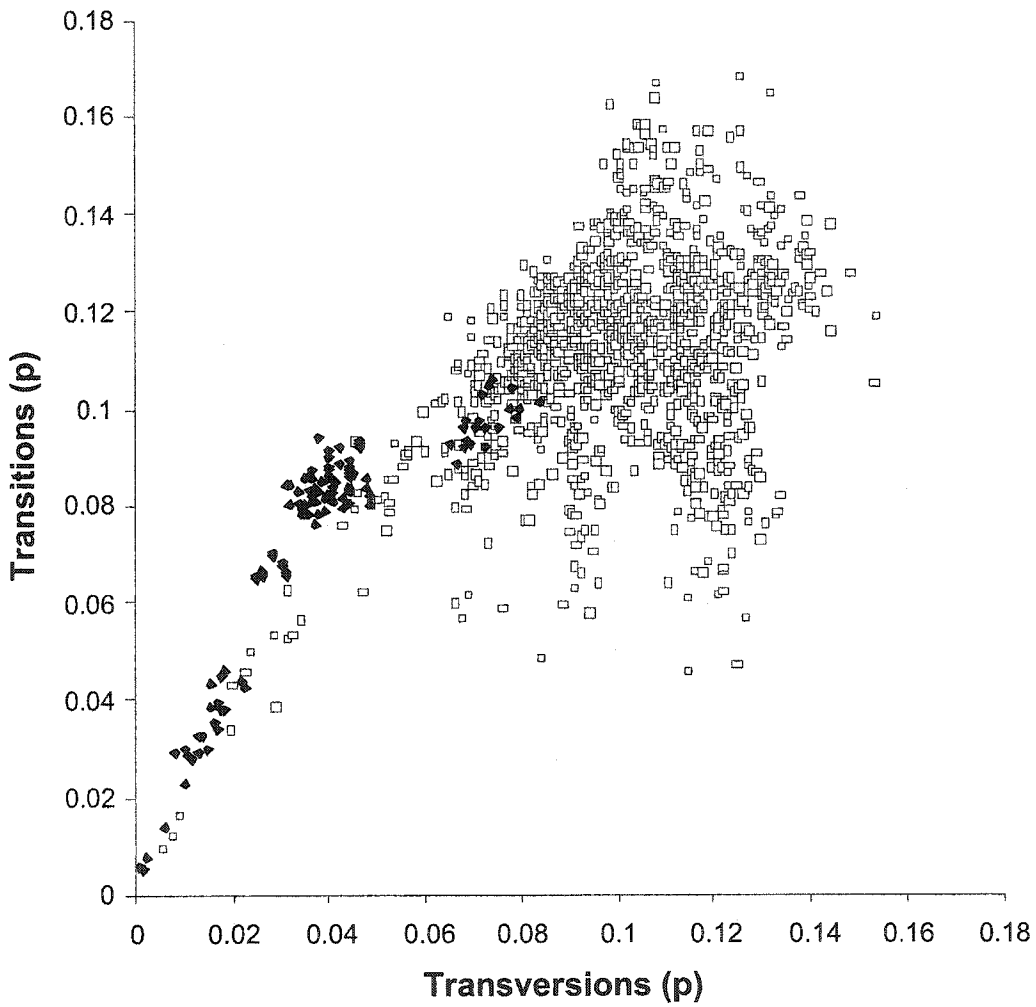
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Transitions (p)
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
Transversions (p)

**Figure 3.2**

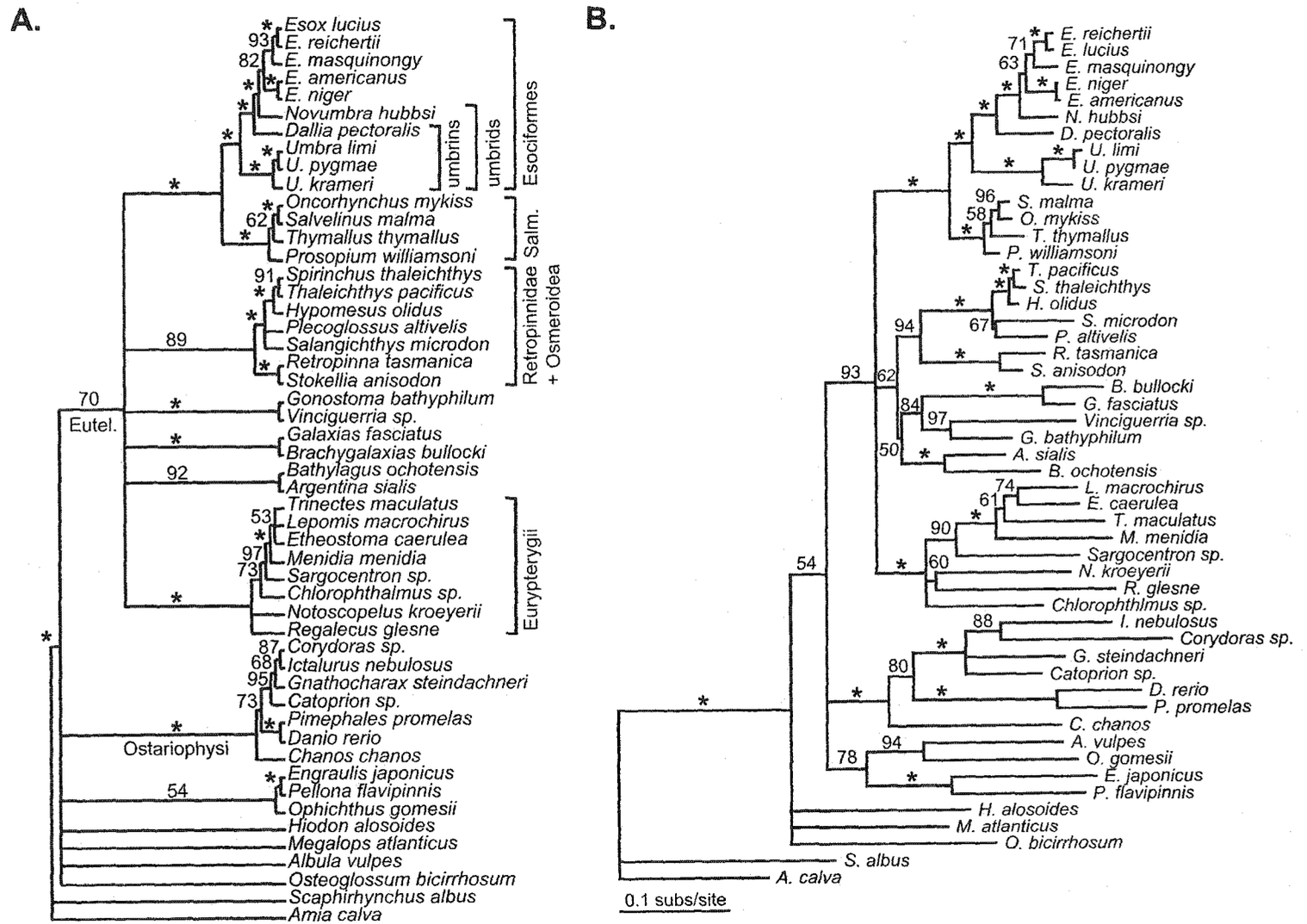
A.
Esox lucius
E. reichertii
E. masquinongy
E. americanus
E. niger
Novumbra hubbsi
Dallia pectoralis
Umbra limi
U. pygmae
U. krameri
umbrins
umbrids
Esociformes
Oncorhynchus mykiss
Salvelinus malma
Thymallus thymallus
Prosopium williamsoni
Salm.
Spirinchus thaleichthys
Thaleichthys pacificus
Hypomesus olidus
Plecoglossus altivelis
Salangichthys microdon
Retropinna tasmanica
Stokellia anisodon
Retropinnidae
+ Osmeroidea
Gonostoma bathyphilum
Vinciguerria sp.
Galaxias fasciatus
Brachygalaxias bullocki
Bathylagus ochotensis
Argentina sialis
Trinectes maculatus
Lepomis macrochirus
Etheostoma caerulea
Menidia menidia
Sargocentron sp.
Chlorophthalmus sp.
Notoscopelus kroeyerii
Regalecus glesne
Eurypterygii
Corydoras sp.
Ictalurus nebulosus
Gnathocharax steindachneri
Catoprion sp.
Pimephales promelas
Danio rerio
Chanos chanos
Engraulis japonicus
Pellona flavipinnis
Ophichthus gomesii
Hiodon alosoides
Megalops atlanticus
Albula vulpes
Osteoglossum bicirrhosum
Scaphirhynchus albus
Amia calva
70
Eutel.
89
92
54
Ostariophysi
B.
E. reichertii
E. lucius
E. masquinongy
E. niger
E. americanus
N. hubbsi
D. pectoralis
U. limi
U. pygmae
U. krameri
S. malma
O. mykiss
T. thymallus
P. williamsoni
T. pacificus
S. thaleichthys
H. olidus
S. microdon
P. altivelis
R. tasmanica
S. anisodon
B. bullocki
G. fasciatus
Vinciguerria sp.
G. bathyphilum
A. sialis
B. ochotensis
L. macrochirus
E. caerulea
T. maculatus
M. menidia
Sargocentron sp.
N. kroeyerii
R. glesne
Chlorophthlmus sp.
I. nebulosus
Corydoras sp.
G. steindachneri
Catoprion sp.
D. rerio
P. promelas
C. chanos
A. vulpes
O. gomesii
E. japonicus
P. flavipinnis
H. alosoides
M. atlanticus
O. bicirrhosum
S. albus
A. calva
0.1 subs/site

**Figure 3.3**

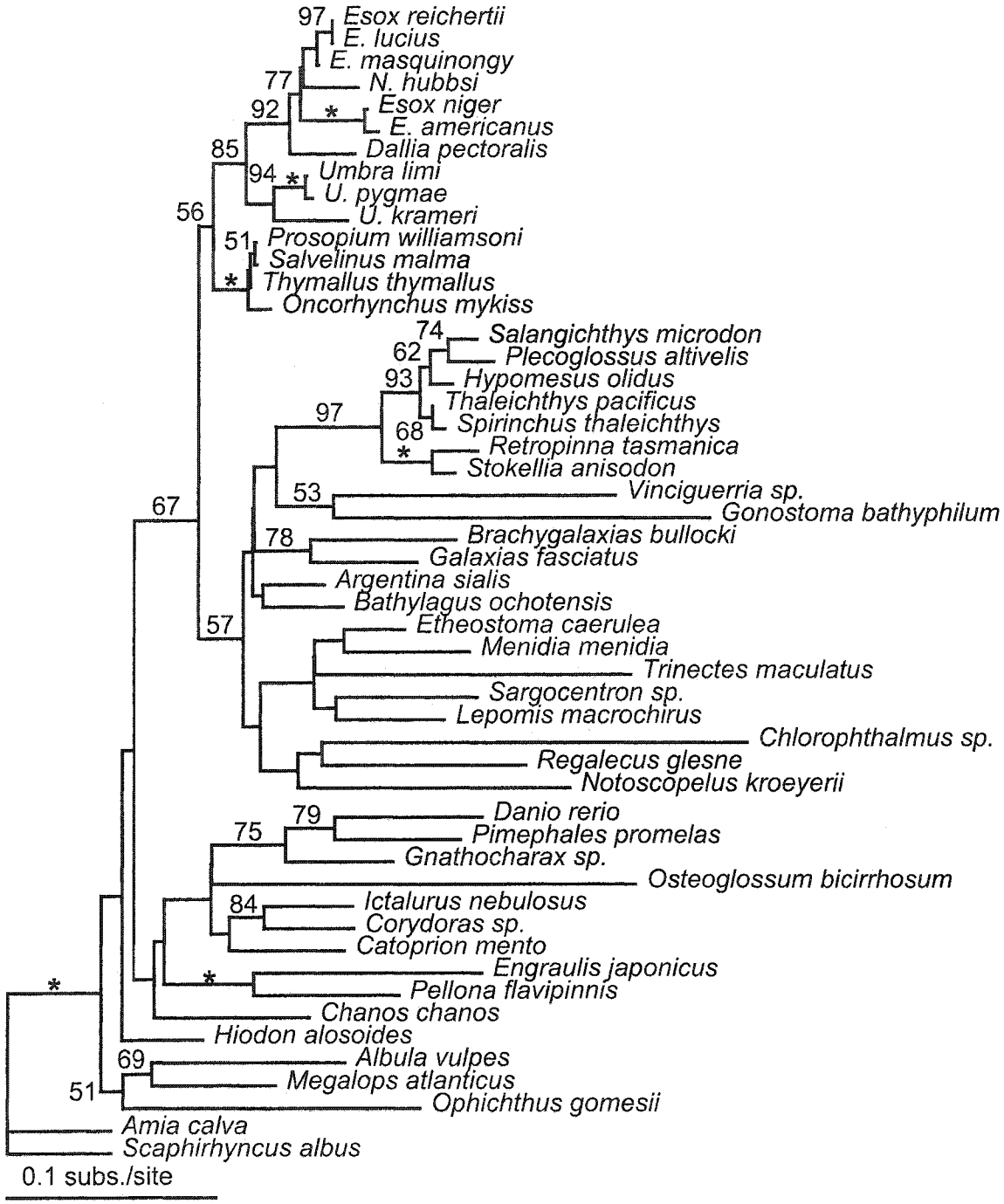
Esox reichertii
E. lucius
E. masquinongy
N. hubbsi
Esox niger
E. americanus
Dallia pectoralis
Umbra limi
U. pygmae
U. krameri
Prosopium williamsoni
Salvelinus malma
Thymallus thymallus
Oncorhynchus mykiss
Salangichthys microdon
Plecoglossus altivelis
Hypomesus olidus
Thaleichthys pacificus
Spirinchus thaleichthys
Retropinna tasmanica
Stokellia anisodon
Vinciguerria sp.
Gonostoma bathyphilum
Brachygalaxias bullocki
Galaxias fasciatus
Argentina sialis
Bathylagus ochotensis
Etheostoma caerulea
Menidia menidia
Trinectes maculatus
Sargocentron sp.
Lepomis macrochirus
Chlorophthalmus sp.
Regalecus glesne
Notoscopelus kroeyerii
Danio rerio
Pimephales promelas
Gnathocharax sp.
Osteoglossum bicirrhosum
Ictalurus nebulosus
Corydoras sp.
Catoprion mento
Engraulis japonicus
Pellona flavipinnis
Chanos chanos
Hiodon alosoides
Albula vulpes
Megalops atlanticus
Ophichthus gomesii
Amia calva
Scaphirhyncus albus
0.1 subs./site

**Figure 3.4**

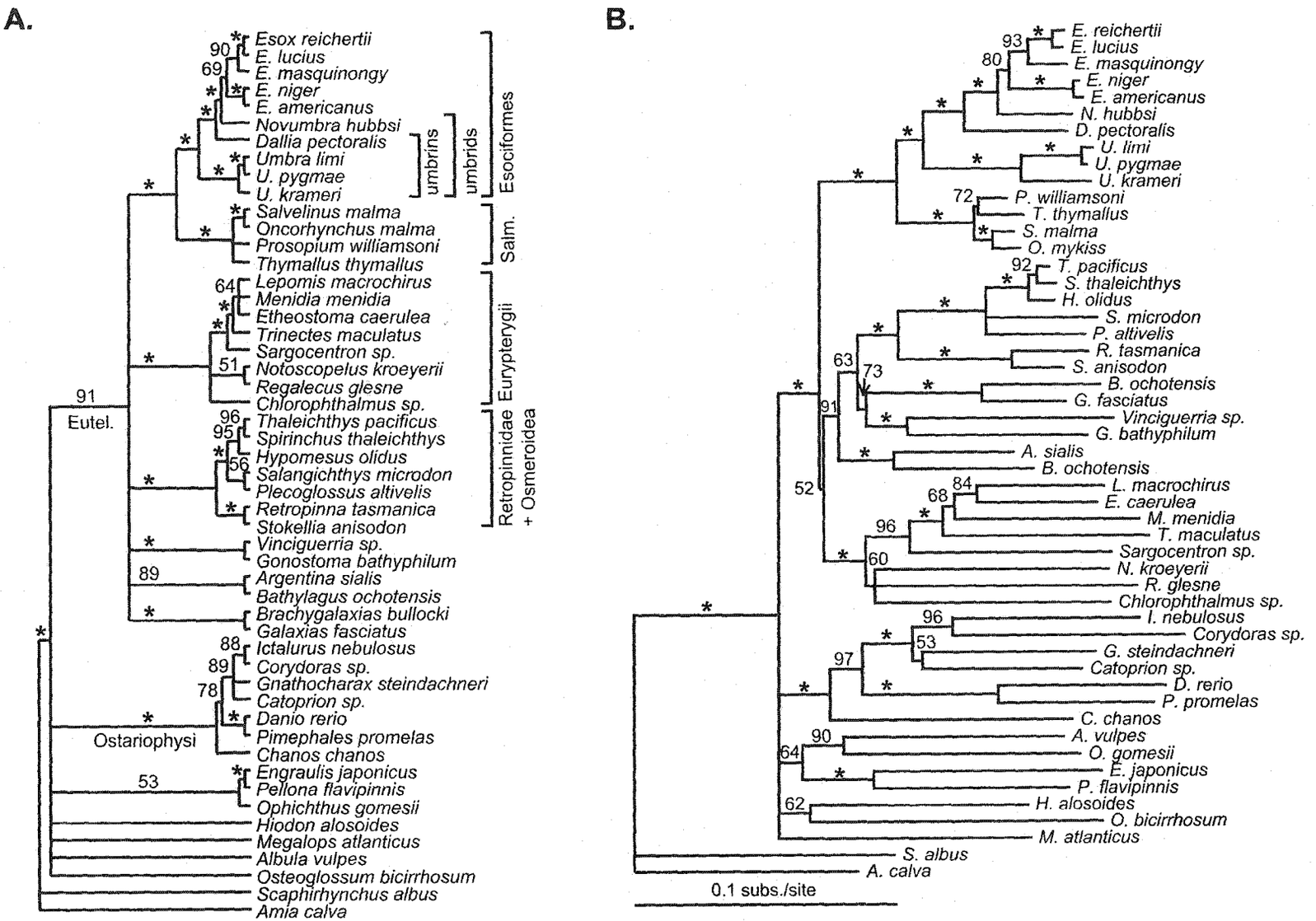
A.
Esox reichertii
E. lucius
E. masquinongy
E. niger
E. americanus
Novumbra hubbsi
Dallia pectoralis
Umbra limi
U. pygmae
U. krameri
Salvelinus malma
Oncorhynchus malma
Prosopium williamsoni
Thymallus thymallus
Lepomis macrochirus
Menidia menidia
Etheostoma caerulea
Trinectes maculatus
Sargocentron sp.
Notoscopelus kroeyerii
Regalecus glesne
Chlorophthalmus sp.
Thaleichthys pacificus
Spirinchus thaleichthys
Hypomesus olidus
Salangichthys microdon
Plecoglossus altivelis
Retropinna tasmanica
Stokellia anisodon
Vinciguerria sp.
Gonostoma bathyphilum
Argentina sialis
Bathylagus ochotensis
Brachygalaxias bullocki
Galaxias fasciatus
Ictalurus nebulosus
Corydoras sp.
Gnathocharax steindachneri
Catoprion sp.
Danio rerio
Pimephales promelas
Chanos chanos
Engraulis japonicus
Pellona flavipinnis
Ophichthus gomesii
Hiodon alosoides
Megalops atlanticus
Albula vulpes
Osteoglossum bicirrhosum
Scaphirhynchus albus
Amia calva
umbrins
umbrids
Esociformes
Salm.
Eurypterygii
Retropinnidae
+ Osmeroidea
Eutel.
Ostariophysi
B.
E. reichertii
E. lucius
E. masquinongy
E. niger
E. americanus
N. hubbsi
D. pectoralis
U. limi
U. pygmae
U. krameri
P. williamsoni
T. thymallus
S. malma
O. mykiss
T. pacificus
S. thaleichthys
H. olidus
S. microdon
P. altivelis
R. tasmanica
S. anisodon
B. ochotensis
G. fasciatus
Vinciguerria sp.
G. bathyphilum
A. sialis
B. ochotensis
L. macrochirus
E. caerulea
M. menidia
T. maculatus
Sargocentron sp.
N. kroeyerii
R. glesne
Chlorophthalmus sp.
I. nebulosus
Corydoras sp.
G. steindachneri
Catoprion sp.
D. rerio
P. promelas
C. chanos
A. vulpes
O. gomesii
E. japonicus
P. flavipinnis
H. alosoides
O. bicirrhosum
M. atlanticus
S. albus
A. calva
0.1 subs./site

**Figure 3.5**

# CHAPTER 4. MUCH MORE DATA AND EXHAUSTIVE SPECIES SAMPLING DOES NOT IMPROVE RESOLUTION OF THE PHYLOGENY OF LAMNIFORM SHARKS

A paper submitted to *Systematic Biology*

Juan Andrés López, Julie A. Ryburn, Olivier Fedrigo, and Gavin J. P. Naylor

## Abstract

We have assembled and analyzed a large set of DNA sequences from all known extant species of the shark order Lamniformes. Our data set includes more than 5,000 nucleotide sites from four protein encoding genes from the mitochondrial (Cytochrome b, NADH-2 and -4) and nuclear genomes (RAG-1). Our phylogenetic analyses using a diverse array of methods of inference and using various schemes of data partitioning and combining show the absence of discernible phylogenetic signal for all but the most trivial groupings of lamniform sharks (e.g. lamniform monophyly, lamnid monophyly). Further, we find evidence of misleading results in the consistent failure of all analyses to support the monophyly of the genera *Alopias* and *Odontaspis*. Examination of sequence characteristics indicated that both the substitution process and their base composition show evidence of heterogeneity that could potentially account for both the trivial and suspect components of the phylogenetic inference. We conclude that there must be aspects of the history of the lamniform lineage and/or the dynamics of molecular evolution of lamniform genomes that have precluded the preservation of evidence from said history in the form of DNA substitutions, at least to the extent that currently devised methods of inference are able to uncover. As a corollary, we suspect that the accumulation of larger DNA sequence data sets

to investigate lamniform phylogeny may lead to robustly supported groupings produced by method bias and with unknown relation to the underlying biology.

## Introduction

Species acquire novel characteristics as they change through time due to evolution and historical accident. A sequence of speciation events in time is expected to result in a hierarchical pattern in the distribution of these characteristic features among species and groups of species. Thus the genealogy of a set of species' lineages may be retrieved directly from measurements of the distribution of features among representatives of those lineages. This simple principle, first formalized into a method by Hennig (1966), underlies cladistic analysis and is the general idea behind all methods that attempt to determine the phylogenetic relationships between biological lineages. The application of this principle to biological data has shown biological reality to be less cleanly patterned than the expectations raised by the simplicity of the underlying idea: extensive homoplasy and conflict between inferences derived from different data sets are common. To be fair, most relevant theoretical work has been silent about the expected prevalence and case-specific variation of historically informative patterns across different levels of biological organization.

Phylogenetic studies have examined features at many different levels of biological organization in search of the historical signal expected to be preserved in the distribution of feature variation. Earlier studies focused on the gross morphology of organisms, while more recently, that focus has shifted to sub-cellular components with a vast emphasis on the sequence of nucleotides in their genomes. Often, the phylogenetic hypotheses supported by different sources of evidence are in conflict, or the support for any given hypothesis is weak. However, given the logical clarity of the theoretical framework and extensive work on its

methodological implementation, it is widely believed that conflicting hypotheses or lack of resolution is caused by the use of insufficient and/or uninformative data. This implies that larger data sets derived from a variety of sources will solve the problems of conflict or lack of resolution found in smaller data sets (Cummings et al., 1995).

As more data are examined, the patterns derived from history are expected to prevail over random noise and patterns derived from biases in the mutation processes. However, this expectation rests on the assumptions that noise is in fact randomly distributed and biases are not concordant across different sources of data (e.g. Naylor and Brown, 1998). These assumptions should be examined for each data set. Another potential problem in phylogenetic studies arises from the dynamics of the speciation process underlying the diversity under study. The history of an explosive radiation will be more difficult to capture than a slower, more uniformly timed process of species generation because rapid radiation may limit the opportunities for synapomorphic (or clade-defining) traits to arise, but all methods of phylogenetic inference are designed to maximize the proportion of the observed variation that can be fitted to a single common history of lineages, thus biases in the methods may reveal clade-defining variation where none is present (e.g. Swofford et al., 2001).

We have assembled a large data set of gene sequences to examine the ability of phylogenetic inference methods to elucidate relationships among sharks in the order Lamniformes and to examine the effects of the interaction between features of the cladogenetic history of the assemblage, the molecular evolution of the DNA sequence characters, and the methods of phylogenetic inference. The order Lamniformes groups 15 extant species of sharks that represent a much-reduced diversity than that thought to have characterized the group in the past (Compagno, 1988; but see Applegate and Espinosa-Arrubarrena, 1996). This impoverished species assemblage exhibits behavioral and

physiological specializations that are interesting to evolutionary biologists. The distribution of these adaptations on a tree depicting lamniform inter-relationships will indicate how often these lamniform adaptations have evolved. In turn, this information combined with studies on the physical bases of the different adaptations will permit discussion of the mode of evolution of complex traits and possibly the convergent evolution of such traits.

The lamniform adaptations that have captured most attention are those of endothermy and planktivory. Both these traits are very rare in elasmobranchs, yet, based on currently accepted hypotheses of phylogeny, each one appears to have originated at least twice within Lamniformes (Compagno, 1988; Morrissey et al., 1997; Martin and Naylor, 1997; Maisey, 1985). Planktivory in elasmobranchs is restricted to three species, two of which can be confidently assigned to Lamniformes: *Cetorhinus maximus*, the basking shark and *Megachasma pelagios*, the megamouth shark. Maisey (1985) proposed a sister group relationship for these species based on their shared planktivory. However, Compagno (1990) examined the anatomical adaptations associated with planktivory in *C. maximus* and *M. pelagios* and found them grossly different, a fact that suggested the multiple origins of that trait. This arrangement is more congruent with Compagno's phylogenetic hypothesis of lamniforms derived from extensive, but not analytically exacting, studies based on morphological evidence (Compagno, 1988, 1990; Figure 4.1A).

Compagno's (1990) hypothesis of lamniform relationships implies either multiple origins or a loss of the anatomical adaptations associated with endothermy. The lamniform families Alopiidae and Lamnidae exhibit endothermy derived from counter-current heat exchange organs of the circulatory system that concentrate metabolic heat in specific organs and tissues. In the currently accepted hypothesis of lamniform phylogeny, lamnids and *Cetorhinus maximus* are sister taxa, and alopiids are the sister group to the lamnid-cetorhinid

clade (Compagno, 1988). Because some reports ascribe rudimentary adaptations for endothermy to *C. maximus* (Carey et al., 1971), Compagno's phylogenetic arrangement of lamniform taxa implies either the parallel refinement of endothermy in species of *Alopias* and the lamnids or the reversal of this trait to a less specialized condition in *C. maximus*. Some differences are evident in the extent and characteristics of the anatomical adaptations associated with endothermy in lamnids and alopiids (Carey et al., 1985; Alexander, 1998). Lamnids show the most highly developed and extensive counter current exchange organs. However, it is difficult to isolate anatomical differences that clearly indicate the separate development of this adaptation in the two lineages. At the same time, more recently advanced phylogenetic hypotheses propose a more distant relationship between lamnids and alopiids, which weakens the hypothesis of single origin of endothermy followed by reversals in the intervening lineages (Martin and Naylor, 1997; Naylor et al., 1997).

The most recent studies of lamniform phylogeny have used DNA sequence data from mitochondrial genes (e.g. Morrissey et al., 1997; Naylor et al., 1997). To date the most complete samples of lamniform taxon diversity have been those of sequences from cytochrome *b* and ND-2 (Naylor et al., 1997). Naylor et al.'s (1997) study is based on approximately 2,300 nucleotides from 14 of the 15 living species of lamniforms, and numerous outgroup species. This large data set does not support a well-resolved set of relationships for the lamniforms. However, some sets of relationships seem to have robust support. The family Lamnidae, as currently defined to include *Lamna*, *Isurus*, and *Carcharodon*, is very strongly supported by tree steps and bootstrap values. Within Lamnidae, there is some support for a sister group relationship between *Isurus* and *Carcharodon*, but this relationship is contradicted in some of the analyses. A sister group relationship between *Cetorhinus maximus* and Lamnidae is also well supported.

Interestingly, there is no support found for the monophyly of Alopiidae, a family that consists of the three species in the genus *Alopias* (the thresher sharks) which are all readily diagnosed by synapomorphies of external anatomy (most strikingly, an elongated upper lobe of the caudal fin) and adaptations associated with endothermy. The genera *Odontaspis*, *Pseudocarcharias*, *Megachasma*, and *Alopias* tend to cluster in an unresolved polytomy. Only one of the two species of *Odontaspis* (the sand tigers) was included in that study. *Mitsukurina owstoni* tends to be assigned a basal position among lamniforms in the majority of analyses. Naylor et al. (1997) summarized their results in a conservative proposal of phylogeny of lamniforms (Figure 4.1B).

In the present study, we used DNA sequences from four genes (three from the mitochondrial genome, and one from the nuclear genome) from specimens representing all fifteen of the living species of lamniform sharks to investigate the phylogenetic relationships among them. Given the weight of the morphological evidence supporting the monophyly of the genera *Alopias* and *Odontaspis*, we judged the merits of inferences produced by different methods by the placement of those species. Specifically, we expected the species of these two genera to form monophyletic groups in any reasonable reconstruction of lamniform relationships. We examine some of the substitution characteristics of the data to determine whether assumptions implied in our methods of phylogenetic inference are justified and to uncover possible reasons for the lack of phylogenetic resolution in our results.

**Research Methods**

To test previously proposed hypotheses of phylogeny among lamniform sharks, we used as characters the nucleotides of the DNA sequences of three mitochondrial genes: Cytochrome *b* (Cyt *b*), NADH-dehydrogenase subunits 2 (ND2) and 4 (ND4); and one single

copy nuclear gene: Recombination activating gene 1 (RAG1). The data set included a total of 5,212 unambiguously aligned nucleotide sites. These include the Cyt *b* and ND2 sequences previously examined by Naylor et al. (1997). All extant lamniform taxa, including *Odontaspis noronhai*, which was not considered in Naylor et al., 1997, were represented in the study. Table 4.1 lists the origin of the lamniform specimens from which DNA sequences were obtained. Four outgroup taxa were included in the phylogenetic analyses: the carcharhiniforms, represented by species of the genera *Carcharhinus* and *Apristurus*, a representative of the orectolobiform genus *Orectolobus*, and one of the heterodontiform genus *Heterodontus*.

Initially, we treated sequences from different genes independently to assess the extent of conflicting patterns between genes. We tested the base composition stationarity for each gene and codon position partition of each gene using a $\chi^2$-test as implemented in PAUP* and evaluated base composition bias differences between lamniform taxa using three-dimensional plots. In these plots, each taxon is represented by a point in a three-dimensional space where the axes represent the percent base composition of three of the four nucleotides. These plots illustrate in an intuitive way the quality of base composition differences detected by the $\chi^2$-test.

We plotted the observed uncorrected proportion of pairwise transitions against transversions to detect substitution saturation or changes in the dynamics of the substitution process. These plots do not allow the distinction between these two possibilities, however both can affect the accuracy of phylogenetic inference. In the absence of base substitution saturation or changes in the substitution process, these plots would show two straight lines diverging at a constant rate. The rate of divergence between the lines is a reflection of the transition:transversion ratio of the underlying base substitutions.

For each of the four sets of gene sequences, we conducted phylogenetic analyses with parsimony, distance and likelihood optimality criteria. In the distance analyses, we determined minimum evolution trees for corrected distance matrices using uncorrected, maximum likelihood corrected and LogDet corrected distances. These last were used to attempt to reduce the effect of base composition heterogeneity on the phylogenetic reconstruction. The settings for likelihood analyses were determined with Modeltest (v. 2.0; Posada and Crandall, 1998), which uses likelihood ratio tests and the Akaike Information Criterion (AIC) to choose among commonly used substitution models that which best fits the data. We tested for homogeneity between different partitions of the data set using the Incongruence Length Difference (ILD) test (Farris et al., 1995) as implemented in PAUP* (Swofford, 1996) and by comparing the estimates of phylogeny derived from each data partition.

To accommodate the some of the heterogeneity between different partitions of the data (e.g. gene, codon position, genome), we conducted Bayesian analyses of the combined data set allowing independence between the model parameters of each partition. These analyses were performed in MrBayes 3.0 (Renquist and Huelsenbeck, 2003). Based partly on the results of the above procedures, we devised a weighting scheme for the combined data set, then we compared bootstrapped MP and ME trees derived from the combined weighted data to previously proposed hypotheses of lamniform interrelationships.

## Results

### *Cytochrome b*

The cytochrome *b* sequences in our analyses included 976 nucleotide sites excluding sites with missing data. The base composition is highly, but not significantly, different

among the taxa examined (P = 0.19777). However, when codon positions are considered separately, third codon position sites show a highly significant deviation from stationarity. All taxa show anti-G bias at the third codon position sites, but the intensity of this bias is weaker in *Isurus*, *Lamna* and *Cetorhinus* than in the remaining lamniform taxa. *Alopias vulpinus*, *Pseudocarcharias kamoharai* and *Odontaspis ferox* show the most extreme anti-G bias, each having only two G's among 320 third codon position examined. Although not statistically significant (P = 1.0), a similar pattern of base composition bias differences can be seen at 1st and 2nd codon position sites, where all lamnid taxa show a higher proportion of G's than the remaining lamniforms. Figure 4.2 shows three-dimensional representations of base composition differences among lamniform taxa for nucleotide sites that deviate from stationarity (i.e. third codon position sites of the mitochondrial genes; Figure 4.2A) and for those that do not differ significantly (Figure 4.2B).

The substitution plots show a clear difference in the transition:transversion ratio (ts:tv) between within-lamniform pairwise comparisons and lamniform-to-outgroup comparisons (Figure 4.3A). This difference is most marked at third codon position sites, not evident at second codon position sites and intermediate at first codon position sites (data not shown).

Using the AIC selection criterion in Modeltest, the model that best fits the base substitution patterns observed in the cytochrome *b* data is the general time reversal model with among site rate variation (GTR+G). Figure 4.4A shows the results of ML analysis with the likelihood settings selected by the Modeltest procedure. Generally, the results of this and other phylogenetic analyses (MP, ME) of cytochrome *b* agree with those reported by Naylor et al. (1997), but there are some differences that may be the result of our choice of outgroup taxa. Bootstrap with uniformly weighted parsimony does not support a monophyletic

Lamniformes because the two outgroup clades form an unresolved polytomy that includes most lamniform taxa. *Odontaspis noronhai*, a species that was not represented in Naylor et al.'s study is not supported as the sister taxon of its congener *O. ferox*. The only clades with strong support are those representing the family Lamnidae (89%) and the genus *Lamna* (100%). There is weak support (50%) for a clade containing *Carcharias*, *Cetorhinus* and the Lamnidae, for the monophyly of *Isurus* (56%) and for the sister group relationship of *Lamna* and *Carcharodon* (50%; Figure 4b). All other lamniform relationships are unresolved in the bootstrap analysis. Weighted parsimony analysis (2:1 ts:tv at third codon position sites) results in marginal support for lamniform monophyly (72%). Within Lamniformes, relationships are the same as those supported by evenly weighted parsimony with the exception of the position of *Carcharodon*, which is weakly supported as the sister group of *Isurus*.

*ND2*

The ND2 sequences included 970 sites excluding sites with missing data. The base composition is significantly different when all the sites are considered ($P = 0.0114$). First and second codon position sites do not deviate from base composition stationarity when tested independently and combined ($P \cong 1.0$). Base composition stationarity is significantly violated ($P << 0.001$) at third codon position sites. The pattern of differences in base composition that were evident in the cytochrome *b* sequences is repeated in the ND2 sequences. At third codon position sites, the lamnid taxa show weaker anti-G bias than the other lamniforms. *Isurus* and *Lamna* also show higher (not significant) proportion of G's at first and second codon position sites; however *Carcharodon* shows similar levels of G's at these sites as the remaining lamniforms. Interestingly, *Alopias superciliosus* shows a

proportion of G's at these sites that is closer to that of *Isurus* and *Lamna* than that of *Carcharodon* and that is markedly different from that of its congeners.

The patterns in ts:tv evident in the substitution plot of ND2 sequences are similar to those observed for cytochrome *b* (Figure 4.3B). The differences in ts:tv between comparisons involving outgroup and ingroup taxa and those restricted to ingroup taxa are more marked in ND2 than in cytochrome *b*. These differences are large for first and third codon position sites and not evident at second codon position sites. Parting from the cytochrome *b* observations, lamniform pairwise comparisons of second codon position sites involving *Mitsukurina* are not clearly different from those among the remaining lamniforms.

Using the AIC selection criterion in Modeltest, the model that best fits the base substitution patterns observed in the ND2 data is the HKY (Hasegawa *et al.*, 1985) with invariable sites and ASRV. Figure 4c shows the results of ML analysis with the settings specified by the Modeltest procedure. Parsimony and distance analyses of ND2 sequences support a monophyletic Lamniformes (58 - 100%). ND2 sequences also give very strong support for the monophyly of the Lamnidae (100%) and strong to marginal support for the monophyly of *Odontaspis* (67 - 95%) under both parsimony and distance optimality criteria. Within the Lamnidae, *Lamna* and *Isurus* are supported as sister genera (60 - 83%). All other lamniform relationships remain unresolved in the majority-rule consensus trees of all analyses under both parsimony and distance optimality criteria (e.g. parsimony; Figure 4.4C).

*ND4*

The ND4 sequences comprised 675 sites excluding missing data. When all sites are considered, lamniform taxa do not significantly deviate from base composition stationarity, but show large differences (P = 0.06654). When the sites are considered separately by codon

position, sites at first and second codon positions show a homogeneous base composition ($P \cong 1.0$) and sites at the third codon position show a significantly heterogeneous base composition ($P << 0.01$). The patterns of base composition of the ND4 sequences among lamniforms are similar to those of the other two mitochondrial genes. Namely, *Isurus* and *Lamna* show weaker anti-G bias at third codon position sites than the remaining taxa. Patterns evident in the ND4 substitution plots are similar to those described above for ND2 with the exception that ingroup to outgroup comparisons of ND4 sequences do not show strong ts:tv differences from those restricted to the ingroup (Figure 4.3C).

Using the AIC selection criterion in Modeltest, the model that best fits the base substitution patterns observed in the ND4 data is the general time reversal model with among site rate variation (ASRV). Figure 4e shows the results of ML analysis with the settings specified by the Modeltest procedure. Overall, results of phylogenetic analyses of ND4 sequences using different optimality criteria support lamniform (63 - 97%) and lamnid (68 - 97%) monophyly. They also give marginal support for *Cetorhinus* as the sister group of the Lamnidae (53 - 76%) and for *Carcharias* as the sister group of *Cetorhinus* and Lamnidae (54 - 68%). In all parsimony analyses, *Isurus* and *Lamna* are supported as sister genera (64 - 75 %), but distance analyses do not resolve relationships among lamnid genera. The genera *Isurus* and *Lamna* are supported as monophyletic, but *Odontaspis* and *Alopias* are not. Uncorrected and HKY85-corrected distances give marginal support to an unresolved clade that includes *Alopias*, *Odontaspis*, *Pseudocarcharias* and *Megachasma*. The majority rule consensus tree of parsimony bootstrap analysis of the ND4 sequences is shown in Figure 4.4F.

*RAG1*

The RAG1 sequences comprised 2800 sites, which when examined both together and separately by codon position exhibit base composition homogeneity among all the taxa included in the study ($P \cong 1.000$). However, three-dimensional plots of base composition show a similar pattern of variation as that observed for the mitochondrial genes (e.g. Figure 4.2B). The lamnid taxa tend to have a higher proportion of G's. This trend is most evident at third codon position sites. Interestingly, the RAG1 gene of *Mitsukurina owstonii* exhibits a very distinct base composition compared to the rest of the lamniforms.

The plot of substitutions for the RAG1 sequences reveals extensive differences between the substitution patterns of these and the mitochondrial gene sequences (Fig. 4.3D). In absolute terms, pairwise comparisons of RAG1 sequences show lower levels of substitutions than those of mitochondrial sequences. The difference is most marked for transitions, which in mitochondrial sequences are observed at proportions six-fold higher than in RAG1 sequences. The difference in the percent transversions among pairwise comparisons of mitochondrial sequences ranges between two and three times the percent transversions observed in RAG1.

All analyses of RAG1 sequences give very strong support to the monophyly of Lamniformes. Strong support is also evident for an early divergence of *Mitsukurina* from the lineage leading to the remaining extant lamniform taxa. A sister species relationship is consistently and robustly supported for *Alopias vulpinus* and *A. pelagicus* (76 - 82%). There is very strong support for the Lamnidae (100%) and, among lamnids, there is weak to no support for a sister group relationship between *Isurus* and *Carcharodon* (0 - 63%). *Cetorhinus* is consistently, but weakly supported as the sister group of the lamnids (50 - 74%). A sister group relationship between this last clade and a clade formed by *Odontaspis*

*noronhai* and *Carcharias taurus* is supported in all the analyses (87 - 96%). The genera *Odontaspis*, *Isurus* and *Alopias* are not found monophyletic in any of the analyses of RAG1 sequences (e.g. Figure 4.4G and H). The model selection procedure indicates that the RAG1 data is best modeled by TrN (Tamura and Nei, 1993) with ASRV. ML analysis using settings empirically determined with Modeltest supports a phylogeny that is in general agreement with those resulting from other analyses (Figure 4.4G).

*Partition Homogeneity*

Two of the six possible pairwise comparisons among genes fail to reject the null hypothesis of homogeneity of phylogenetic structure, as determined by the ILD test (Farris et al., 1995) implemented in PAUP* (Table 4.2). The ILD tests of the pairs involving ND2 and RAG1, and ND2 and Cyt *b* result in rejection of the null hypothesis of homogeneity ($\alpha$ = 0.01; Cunningham, 1997; but see Barker and Lutzoni, 2002). The remaining gene pairwise comparisons fail to reject the hypothesis of homogeneity. Phylogenetic structure of the grouped mitochondrial genes and RAG1 are significantly different. No pairwise test between codon position partitions of the entire data set is significant. However, when individual genes are divided by codon position and these subsets are compared between different genes a pattern emerges where phylogenetic structure at third codon positions is more heterogeneous than at other codon positions and comparisons of first codon positions consistently result in the highest p-values.

*Combined Data Set*

The entire data set fails to produce a fully resolved phylogeny under different criteria of optimality, different character weighting schemes, and different models of sequence evolution. More significantly, no method resulted in a phylogenetic reconstruction where

*Alopias* and *Odontaspis* are both supported as monophyletic genera. The family Lamnidae is well supported in all analyses, as is the basal position of *Mitsukurina owstoni* among the lamniforms. The lamniforms minus *M. owstoni* are resolved into two large clades: one containing the family Lamnidae, *Cetorhinus maximus* and *Carcharias taurus*; and the other containing the Alopiidae, the Odontaspidae, *Megachasma pelagios* and *Pseudocarcharias kamoharai* (as found by Naylor et al., 1997). Figure 4.5 shows the tree topology of the consensus tree of 10000 bootstrap pseudoreplicates of the combined sequence data using the parsimony optimality criterion with heuristic searches for the most parsimonious tree for each pseudoreplicate.

**Discussion**

Our results mirror the lack of resolution seen in previous molecular studies and conflict with the published hypothesis of relationship based on morphological character data presented by Compagno (1990). Some might argue that the morphological hypothesis, being based on a smaller character data set, should be given less credence than the molecular based hypothesis presented herein (Russo et al., 1996). However we take a different view and maintain that there are reasons to also be suspicious of the molecular inference. The most compelling cautionary flag raised by the molecular data is that the three species of thresher shark do not fall out as a monophyletic group. While it is possible that the three species represent the outcome of morphological convergence, parallelism, or differential retention of ancestral conditions, it would be so spectacular a case as to stretch the bounds of credulity. Similarly the two species of sandtiger *Odontaspis taurus* and *O. noronhai*, do not fall out as each other's closest living relatives. We regard these groupings as phylogenetic positive controls. The fact that such obvious natural groups do not form monophyletic groups, in our

view, casts doubt on all of the inferences suggested by standard analyses of the molecular tree. The only relationships seen in the molecular tree that are both strongly supported and consistent with morphological data are: the monophyly of the genus *Isurus*; the monophyly of the genus *Lamna*; the monophyly of the family Lamnidae and the placement of *Mitsukurina* as the most basal extant member of the group.

Inadequate taxon-sampling and insufficient data are perhaps the two most commonly cited "explanations" given to account for phylogenetic inferences that lack credibility. However, in our case we collected DNA sequence data for 4 different gene fragments, a total of more than 5 kilobases, for all the extant representatives of the clade, and yet still failed to obtain a robust, credible inference. Improved taxon-sampling is simply not possible, as all of the known extant species are already included in the study. It is, of course, possible to collect more character data. However one wonders just how much might be required considering that the more than doubling of the character matrix represented by the data presented here has not translated into any improvements in phylogenetic resolution. More importantly, there is always the concern that collecting more character data will increase the chance of converging to an incorrect inference that is well supported due to statistical inconsistency in the estimation method used (Felsenstein, 1978; Huelsenbeck and Hillis, 1993). Phylogenetic inconsistency arises when the model used to infer the trees is not coincident with the evolutionary process that generated the data (e.g. Takezaki and Gojobori, 1999). The model needs to be unbiased with respect to the value of the parameter being estimated, in this case the tree. An unbiased method will converge on the correct tree if given enough data. By contrast a biased method will converge on an incorrect tree.

To devise improved phylogenetic methods, it is prerequisite to determine the characteristics of the evolutionary process that lead to mis- or uninformative inferences given

current approaches. The molecular evolution process of DNA sequences can be characterized by, among other measures, base composition changes, codon usage differences among and between genes, relative contribution of the different base substitution types, and differences in rates of change between taxonomic units and between genes/genomes. We have examined these characteristics in the characters in our data set in an attempt to determine likely sources of misleading phylogenetic signal.

*Base composition*

The most commonly used method for examining changes in base composition is the chi-square test as implemented in PAUP. This test detects statistically significant deviations in base composition among a set of taxa without considering the underlying phylogenetic structure. The three-dimensional plots we present (Figure 4.2) and examined in the context of the phylogenetic hypotheses allow us to intuitively visualize the effect that changes in base composition among taxa may have on the resulting phylogeny.

All of the plots show that the lamnids and to a lesser extent *Cetorhinus* and *Carcharias*, consistently occupy, as a group, regions of the three-dimensional base space that set them apart from the remaining taxa (Figure 4.2). This pattern persists when subsets of the data that do not significantly deviate from base composition stationarity (i.e. first and second codon position of the mitochondrial genes and all of the RAG-1 sequence) are plotted (Figure 4.2B). Of course, the points in this last plot span a much reduced volume in the base space when compared to that filled by the points in plots of nucleotide sites that do deviate from base composition stationarity. The unique base composition characteristics of these taxa implied by this observation may be responsible for the very robust support for the Lamnidae

and the consistent support for a lamnid-*Cetorhinus-Carcharias* clade, a group that is not compatible with any of the morphology based phylogenetic reconstructions.

Notably, while the points that represent the species of *Lamna* and of *Isurus* in the base composition plots tend to be in close proximity to each other, the points that represent the species of *Alopias* are more dispersed and, for all subsets of the data set, are found interspersed among the other lamniform species. To a lesser extent, a similar pattern is evident in the placement of points representing species of *Odontaspis*.

Considering the effects of base composition differences on the resulting phylogeny, we contemplate two alternatives. Differences in base composition may be the result of synapomorphic changes in substitution bias in an evolutionary lineage and therefore, the phylogenetic inference may be ultimately correct, although technically based on homoplasious support (i.e. right answer, wrong reasons). Strong support for the Lamnidae may be due in part to this phenomenon. Alternatively, base composition bias as an indicator of a restricted evolutionary space may implicate unspecified biological constraints that prevent the appearance and preservation of real synapomorphic character states. We suspect that the reasons for the lack of support for a monophyletic *Alopias* and *Odontaspis* fall under this category.

Table 4.3 shows a comparison of different measures of base composition at synonymous sites of the four genes examined here. The trends seen in the base composition plot are mirrored in the variation seen in the use of certain bases at synonymous sites and, consequently, measures of codon usage such as effective number of codons (Nc; Wright, 1990). These measures for species of *Alopias* and of *Odontaspis* show varying patterns between genes and substantial intergeneric variation relative to the total variation observed

among lamniforms. The variation we observe in these indices of codon usage may be a reflection of differential constraints affecting the molecular evolution of lamniform genomes.

*Transition : Transversion ratio and saturation*

Due in large part to computational costs, current methods of phylogenetic inference assume homogeneity of the evolution process through time even when some may accommodate heterogeneity among aspects and subsets of the character set (e.g. ASRV, substitution matrix parameterization). However, the limited number of character states available to a given nucleotide site, dictate that even a homogenous substitution process will result in apparent substitution heterogeneity, a phenomenon usually termed substitution saturation. Under the best conditions (e.g. random distribution of homoplasious character states among taxa), a data matrix with apparent substitution heterogeneity may be expected to be uninformative concerning the deeper nodes of the phylogeny being reconstructed. But if homoplasy is not randomly distributed (e.g. as a result of differential bias in base composition or the substitution process), the lack of informative characters may be substituted by misinformation if, as explained above, the deviation from randomness is not itself synapomorphic.

Plotting transitions against transversions (Figure 4.3) allowed us to conduct a cursory examination of the apparent characteristics of the substitution spectrum among the taxa in the study for the four genes examined. Of the four genes examined, RAG-1 shows the most uniform relationship between transitions and transversions. This uniform relationship extends to comparisons involving outgroup taxa. On the other hand, ND2 sequences show a drastic departure from uniformity that neatly divides the points in the plot between those resulting from within-ingroup comparisons and the rest. The observations derived from these

plots are not unexpected and they can be explained by differences in the magnitude of divergence reached by RAG-1 and the mitochondrial genes.

Considering the effects that saturation may have had in our analyses, it is worth noting that strong substitution saturation is apparent only in pairwise comparisons that involve at least one outgroup taxon. Further, even in the absence of apparent saturation, RAG-1 sequences are uninformative about the same lamniform relationships not resolved with mitochondrial data. Thus, if substitution saturation is advanced as the cause of the lack of resolution in the phylogenetic inference derived from the mitochondrial sequences, a different argument must be offered to explain the same aspect of the results of analyses of RAG-1.

*Internal and terminal branch lengths*

The phylograms resulting from all the analyses shared long terminal branches and relatively short internodal branches (Figure 4.4). This branch length pattern common to all four gene sequences analyzed indicates two alternative scenarios. If the relative lengths of the branches on the phylograms reflect real differences in the time elapsed between and since the formation of new lineages then extant lamniform diversity is the result of a few bursts of cladogenesis in the distant past. And, our inability to confidently resolve the phylogenetic relationships within the order is the result of an expectedly low accumulation of synapomorphic traits during the short inter-cladogenetic intervals. The alternative scenario holds that the contrasting branch length patterns seen in the phylograms are not reflective of the history of the assemblage, but instead are artifacts produced by the data and the analytical methods employed, by the extinction history of lamniforms or by a combination of these factors. Unfortunately, biologists are often faced with this uncertainty and the facts needed

to determine which scenario is correct are often inaccessible to us. Computations that simulate by approximation some elements of this problem may be devised and used to help determine the level of confidence with which we advance hypotheses of evolutionary relationships. This approach is beyond the scope of this report and will be discussed elsewhere.

*Lamniform phylogeny*

The results of our phylogenetic analyses agree with part of the hypothesis offered by Naylor *et al.* (1997). The result of bootstrap analysis of the combined data set strongly supports: the basal position of *Mitsukurina*, the monophyly of the Lamnidae, and the sister group relationship between *Cetorhinus* and the Lamnidae (Figure 4.5). In addition, and differing from previously proposed hypotheses, our analyses yield moderate to robust support for a sister group relationship between *Carcharias* and the *Cetorhinus*+Lamnidae clade. This relationship is consistently supported by different types of analyses of different subsets of the data set. Support for the monophyly of the genus *Odontaspis* is not robust and it is only found in analyses of the ND2 sequence data. Our data and analyses do not support the monophyly of the genus *Alopias*. We refrain from advancing a new hypothesis of lamniform phylogeny considering that the results of our analyses do not seem to represent an improvement on or a departure from current hypotheses.

If we are to improve our understanding of lamniform phylogeny using molecular data, how do we proceed? There are two approaches we can take: Find a model that more accurately describes the data, or identify and remove subsets of the data that appear to violate the assumptions of the model such that the model is rendered consistent with the remaining data. For this contribution we employed a representative diversity of commonly used models

of phylogenetic inference in to see the effect of the method and model on the inference. We conclude from this study, that there are aspects of the underlying cladogenetic history of living lamniforms or characteristics of the molecular evolution of their genomes that prevent currently devised methods of phylogenetic reconstruction from producing fully resolved and robustly supported inferences of lamniform phylogeny based on samples of character states at nucleotide sites from living lamniform diversity.

**Acknowledgements**

We are very thankful to the NSF for support to GPJN for this research and to the Ecology and Evolutionary Biology program and the Department of Zoology and Genetics at Iowa State University for research and travel support to JAL.

**References**

Alexander, R. L. 1998. Blood supply to the eyes and brain of lamniform sharks (Lamniformes). J. Zool., Lond. 245:363-369.

Applegate, S. P., and L. Espinosa-Arrubarrena. 1996. The fossil history of *Carcharodon* and its possible ancestor, *Cretolamna*: a study in tooth identification. *In* The Biology of the White Shark (P. Klimley, and Ainsley, eds.). Academic Press, New York.

Barker, F. K., and F. M. Lutzoni. 2002. The utility of incongruence length difference test. Syst. Biol. 51:625-637.

Carey, F. G., J. G. Casey, H. L. Pratt, D. Urquhart, and J. E. McCosker. 1985. Temperature heat production and heat exchange in lamnid sharks. Mem So Cal Acad Sci 9:92-108.

Carey, F. G., J. M. Teal, J. W. Kanwisher, and K. D. Lawson. 1971. Warm-bodied fish. Am. Zool. 11:137-145.

Compagno, L. J. V. 1988. Sharks of the order Carcharhiniformes. Pages xxii, 486 , 93 of plates Princeton University Press, Princeton, N.J.

Compagno, L. J. V. 1990. Relationships of the megamouth shark, *Megachasma pelagios* (Lamniformes: Megachasmidae). with comments on its feeding habits. Pages 357-379 *in* Elasmobranchs as living resources: advances in the biology, ecology, systematics, and the status of the fisheries (H. L. Pratt Jr., G. S.H., and T. Taniuchi, eds.). NOAA.

Cummings, M. P., S. P. Otto, and J. Wakeley. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Molecular Biology and Evolution 12:814-822.

Cunningham, C. W. 1997. Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. 14:733-740.

Farris, J. S., M. Kallersjo, A. G. Kluge, and C. Bult. 1995. Constructing a significance test for incongruence. Syst. Biol. 51:19-31.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:27-33.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21:160-174.

Hennig, W. 1966. Phylogenetic systematics. University of Illinois, Urbana.

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247-264.

Maisey, J. G. 1985. Relationships of the megamouth shark, *Megachasma*. Copeia 1985:228-231.

Martin, A. P., and G. J. P. Naylor. 1997. Independent origins of filter-feeding in megamouth and basking sharks (Order Lamniformes) inferred from phylogenetic analysis of cytochrome *b* gene sequences. Pages 39-50 *in* Biology of the megamouth shark (K.

Yano, J. F. Morrissey, Y. Yabumoto, and K. Nakaya, eds.). Tokai University Press, Tokyo.

Morrissey, J. F., K. A. Dunn, and F. Mule. 1997. The phylogenetic position of *Megachasma pelagios* inferred from mtDNA sequence data. Pages 33-37 *in* Biology of the megamouth shark (K. Yano, J. F. Morrissey, Y. Yabumoto, and K. Nakaya, eds.). Tokai University Press, Tokyo.

Naylor, G. J. P., and W. M. Brown. 1998. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst. Biol. 47:61-76.

Naylor, G. J. P., A. P. Martin, E. G. Mattison, and W. M. Brown. 1997. Interrelationships of lamniform sharks: testing phylogenetic hypotheses with sequence data. Pages 199-218 *in* Molecular Systematics of Fishes Academic Press.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Renquist, F., and J. P. Huelsenbeck. 2003.. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Russo, C. A., N. Takezaki, and M. Nei. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13:525-536.

Swofford, D. L. 1996. PAUP: Phylogenetic analysis using parsimony, version 4.0. Sinauer Associates.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Systematic Biology 40:525-539.

Takezaki, N., and T. Gojobori. 1999. Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. Molecular biology and evolution. 16:590-601.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512-526.

Wright, F. 1990. The effective number of codons used in a gene. Gene 87:23-29.

**Table 4.1** Origin of lamniform specimens from which the sequences in this study were obtained

| Species | Specimen Origin |
|---|---|
| *Alopias pelagicus* | Taiwan |
| *A. pelagicus* | Mexico (Pacific) |
| *A. superciliosus* | U.S.A. |
| *A. superciliosus* | U.S.A. |
| *A. vulpinus* | Japan |
| *A. vulpinus* | U.S.A. |
| *Cetorhinus maximus* | England |
| *C. maximus* | England |
| *Carcharodon carcharias* | S. Africa |
| *C. carcharias* | U.S.A. |
| *Isurus oxyrhinchus* | Mexico (Pacific) |
| *I. oxyrhinchus* | U.S.A. |
| *I. paucus* | U.S.A. |
| *Lamna ditropis* | U.S.A.(Pacific) |
| *L. ditropis* | Japan |
| *L. nasus* | U.S.A. |
| *L. nasus* | U.S.A. |
| *Megachasma pelagios* | Japan |

**Table 4.1** (continued)

| Species | Specimen Origin |
|---|---|
| *Mitsukurina owstoni* | Tasmania |
| *M. owstoni* | Japan |
| *Carcharias taurus* | U.S.A. |
| *Odontaspis ferox* | Azores |
| *O. ferox* | Azores |
| *O. noronhai* | Brazil |
| *Pseudocarcharias kamoharai* | Taiwan |
| *P. kamoharai* | Taiwan |

**Table 4.2** The ILD test statistic for comparisons of various partitions of the sequence data.

| | Cyt b | ND2 | ND4 | RAG1 |
|---|---|---|---|---|
| Cytochrome b | | 0.007 [1] | 0.370 | 0.046 |
| ND2 | 0.052 [2] | | 0.236 | 0.001 |
| ND4 | 0.215 | 0.095 | | 0.029 |
| RAG1 | 0.074 | 0.001 | 0.001 | |
| | All sites | 1st | 2nd | 3rd |
| mtDNA vs. RAG1 | 0.029 | 0.997 | 0.708 | 0.047 |

[1] Values above the diagonal are from pairwise tests considering all codon positions.

[2] Values below the diagonal are from pairwise tests considering only third codon position sites.

**Table 4.3** Proportion of G's at third codon position sites and indices of codon usage bias of the lamniform sequences examined in this study.

| Species[1] | Cytochrome b | | | ND2 | | | ND4 | | | RAG1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | G-3[2] | GC3s[3] | Nc[4] | G-3 | GC3s | Nc | G-3 | GC3s | Nc | G-3 | GC3s | Nc |
| Ape | 0.012 | 0.405 | 40.482 | 0.023 | 0.303 | 42.090 | 0.013 | 0.280 | 40.226 | 0.214 | 0.409 | 54.618 |
| Asu | 0.020 | 0.416 | 43.103 | 0.026 | 0.383 | 49.163 | 0.044 | 0.324 | 43.655 | 0.212 | 0.406 | 54.506 |
| Avu | 0.006 | 0.355 | 41.712 | 0.026 | 0.326 | 48.971 | 0.009 | 0.262 | 41.141 | 0.207 | 0.403 | 54.289 |
| Pka | 0.006 | 0.332 | 42.023 | 0.012 | 0.274 | 38.876 | 0.022 | 0.302 | 42.075 | 0.215 | 0.410 | 54.333 |
| Ofe | 0.006 | 0.295 | 43.181 | 0.020 | 0.297 | 41.156 | 0.027 | 0.244 | 39.304 | 0.214 | 0.409 | 54.274 |
| Ono | 0.020 | 0.370 | 45.618 | 0.017 | 0.294 | 46.277 | 0.013 | 0.213 | 39.130 | 0.210 | 0.407 | 54.252 |
| Cta | 0.020 | 0.425 | 44.000 | 0.017 | 0.349 | 44.690 | 0.018 | 0.333 | 41.176 | 0.209 | 0.400 | 54.097 |
| Mpe | 0.023 | 0.335 | 41.232 | 0.035 | 0.297 | 43.072 | 0.022 | 0.249 | 42.170 | 0.209 | 0.404 | 53.950 |

[1] Species abbreviations are as given in caption of Figure 4.1.

[2] Proportion of guanosine at third codon position sites for a given gene.

[3] Proportion of guanosine or cytosine at the 3rd position of synonymous codons.

[4] Effective number of codons: No bias in codon usage is indicated by Nc values of 60 or 61 for vertebrate mitochondrial or universal genetic codes, respectively (Wright, 1990).

**Table 4.3** (continued)

| Species | Cytochrome b | | | ND2 | | | ND4 | | | RAG1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | G-3 | GC3s | Nc | G-3 | GC3s | Nc | G-3 | GC3s | Nc | G-3 | GC3s | Nc |
| Mow | 0.017 | 0.416 | 44.601 | 0.029 | 0.352 | 42.704 | 0.040 | 0.342 | 45.761 | 0.205 | 0.388 | 53.583 |
| Cma | 0.035 | 0.491 | 47.284 | 0.026 | 0.369 | 40.440 | 0.053 | 0.413 | 41.541 | 0.213 | 0.411 | 53.883 |
| Lna | 0.032 | 0.460 | 44.919 | 0.052 | 0.470 | 39.803 | 0.076 | 0.404 | 47.864 | 0.222 | 0.415 | 54.556 |
| Ldi | 0.038 | 0.486 | 43.850 | 0.050 | 0.444 | 41.109 | 0.084 | 0.418 | 46.403 | 0.219 | 0.413 | 54.692 |
| Cca | 0.017 | 0.465 | 46.380 | 0.032 | 0.380 | 45.318 | 0.044 | 0.418 | 40.500 | 0.217 | 0.412 | 54.338 |
| Ipa | 0.040 | 0.526 | 44.716 | 0.058 | 0.458 | 41.793 | 0.076 | 0.476 | 52.180 | 0.216 | 0.410 | 54.307 |
| Iox | 0.029 | 0.468 | 46.720 | 0.074 | 0.461 | 43.606 | 0.080 | 0.462 | 46.885 | 0.219 | 0.420 | 54.977 |

**Figure Captions**

**Figure 4.1** (a) Compagno's (1990) morphology-based hypothesis of lamniform relationships. (b) Naylor et al.'s (1997) mtDNA-based hypothesis of lamniform relationships. Species names are abbreviated as follows: *Alopias pelagicus* to Ape; *A. superciliosus* to Asu; *A. vulpinus* to Avu; *Carcharias taurus* to Cta; *Carcharodon carcharias* to Cca; *Cetorhinus maximus* to Cma; *Isurus paucus* to Ipa; *I. oxyrhyncus* to Iox; *Lamna ditropis* to Ldi; *L. nasus* to Lna; *Megachasma pelagios* to Mpe; *Mitsukurina owstoni* to Mow; *Odontaspis ferox* to Ofe; *O. noronhai* to Ono; *Pseudocarcharias kamoharai* to Pka.

**Figure 4.2** (A) Plot of base composition at third codon position sites from the mtDNA sequences. (B) Plot of base composition from the RAG-1 sequences and the first and second codon position sites from the mtDNA sequences. Note the difference in range magnitude between the plots. Species names are abbreviated as in Figure 4.1.

**Figure 4.3** Plots of observed transitions (o) and transversions (x) versus total observed substitutions for all pairwise comparisons for: (a) Cytochrome *b*, (b) ND2, (c) ND4, (d) RAG1.

**Figure 4.4** Maximum likelihood tree (MLT) using settings selected with the use of Modeltest and majority-rule consensus tree of 10000 bootstrap maximum parsimony pseudoreplicates (BMP) with ML-inferred branch lengths. Inferences from cytochrome *b* (a) MLT, (b) BMP; ND2 (c) MLT, (d) BMP; ND4 (e) MLT, (f) BMP; RAG1 (g) MLT, (h) BMP. Species names are abbreviated as in Figure 4.1

**Figure 4.5** Majority-rule consensus tree of 100000 maximum parsimony pseudoreplicates of the combined data set. Species names are abbreviated as in Figure 4.1.

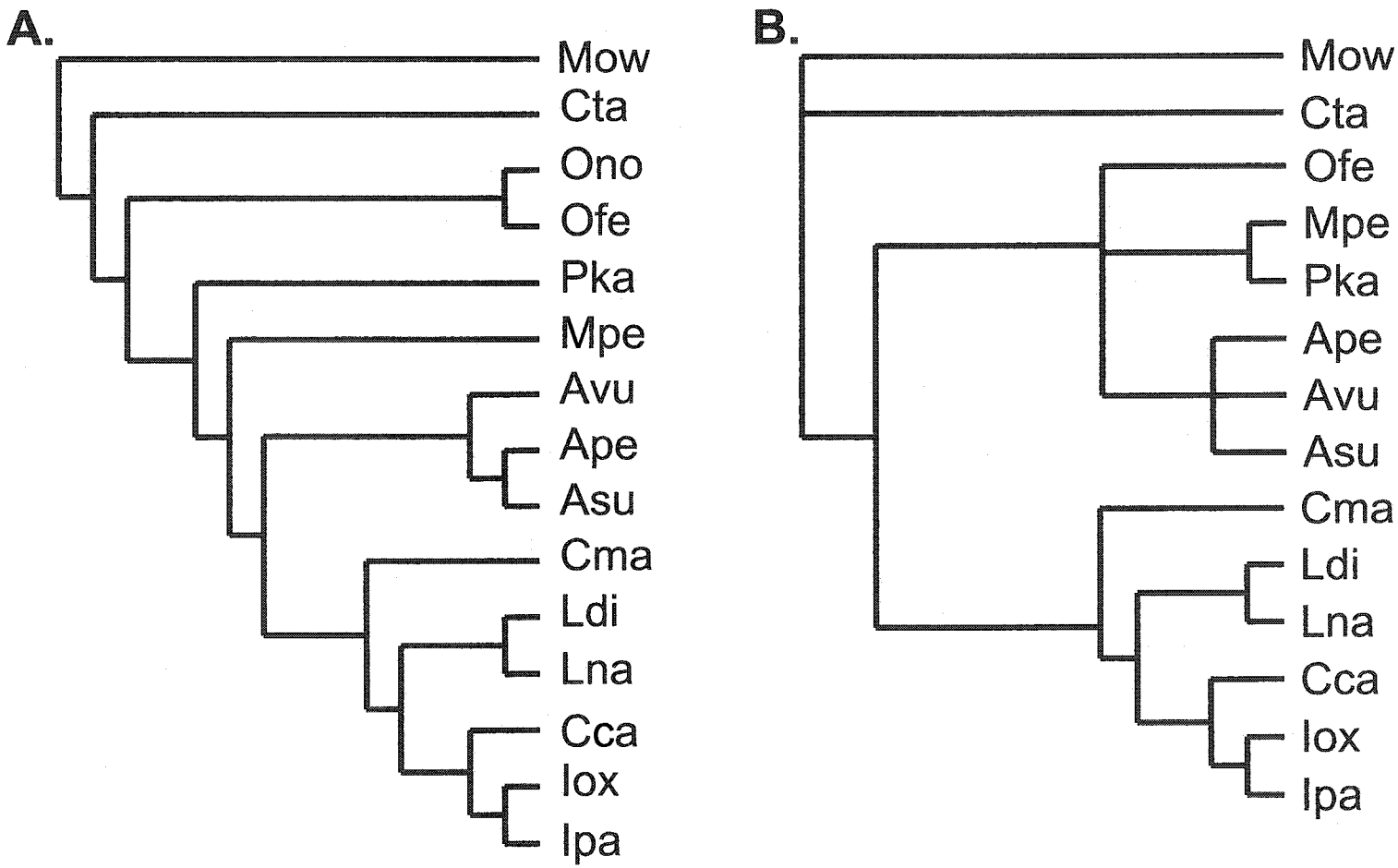
A.
Mow
Cta
Ono
Ofe
Pka
Mpe
Avu
Ape
Asu
Cma
Ldi
Lna
Cca
Iox
Ipa
B.
Mow
Cta
Ofe
Mpe
Pka
Ape
Avu
Asu
Cma
Ldi
Lna
Cca
Iox
Ipa

**Figure 4.1**

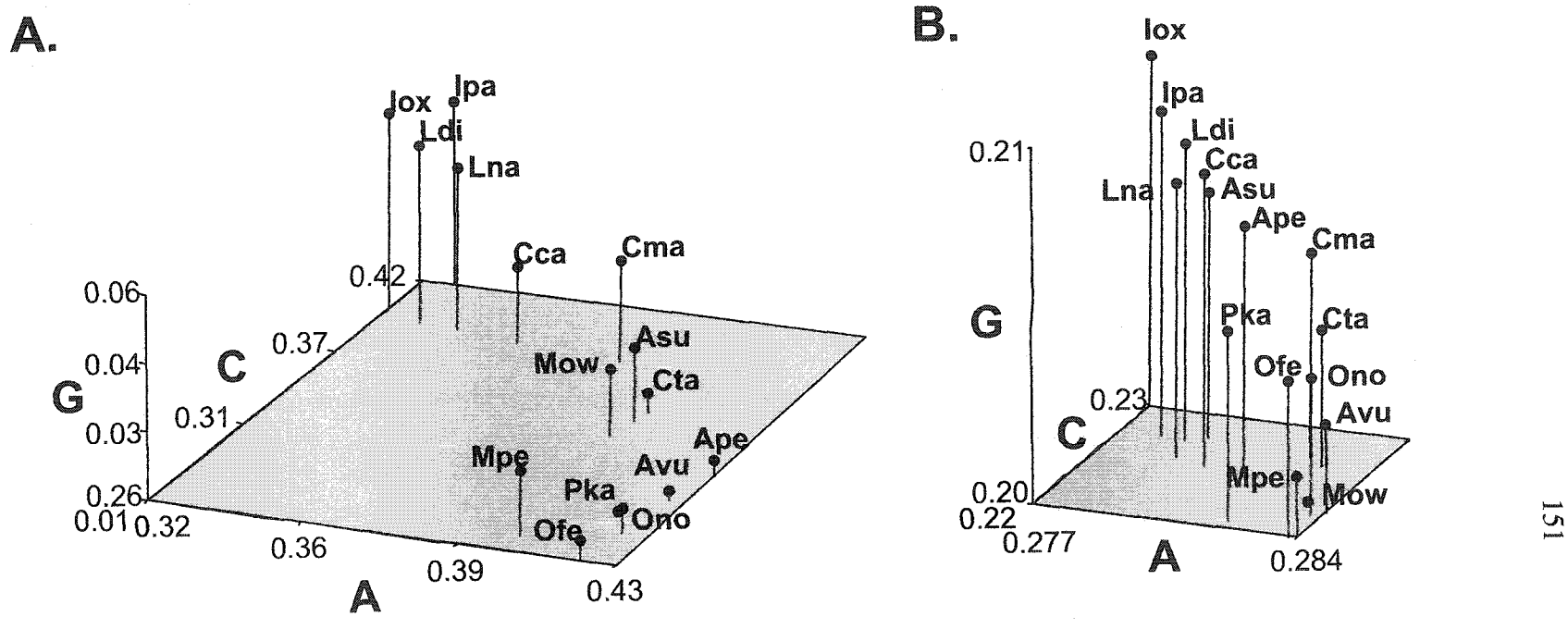
A.
Iox
Ipa
Ldi
Lna
Cca
Cma
Asu
Mow
Cta
Ape
Mpe
Avu
Pka
Ono
Ofe
0.06
0.04
0.03
0.26
G
0.01
0.32
0.36
0.39
0.43
A
0.42
0.37
0.31
C
B.
Iox
Ipa
Ldi
Cca
Asu
Ape
Cma
Lna
Pka
Cta
Ofe
Ono
Avu
Mpe
Mow
0.21
G
0.23
C
0.20
0.22
0.277
A
0.284

**Figure 4.2**

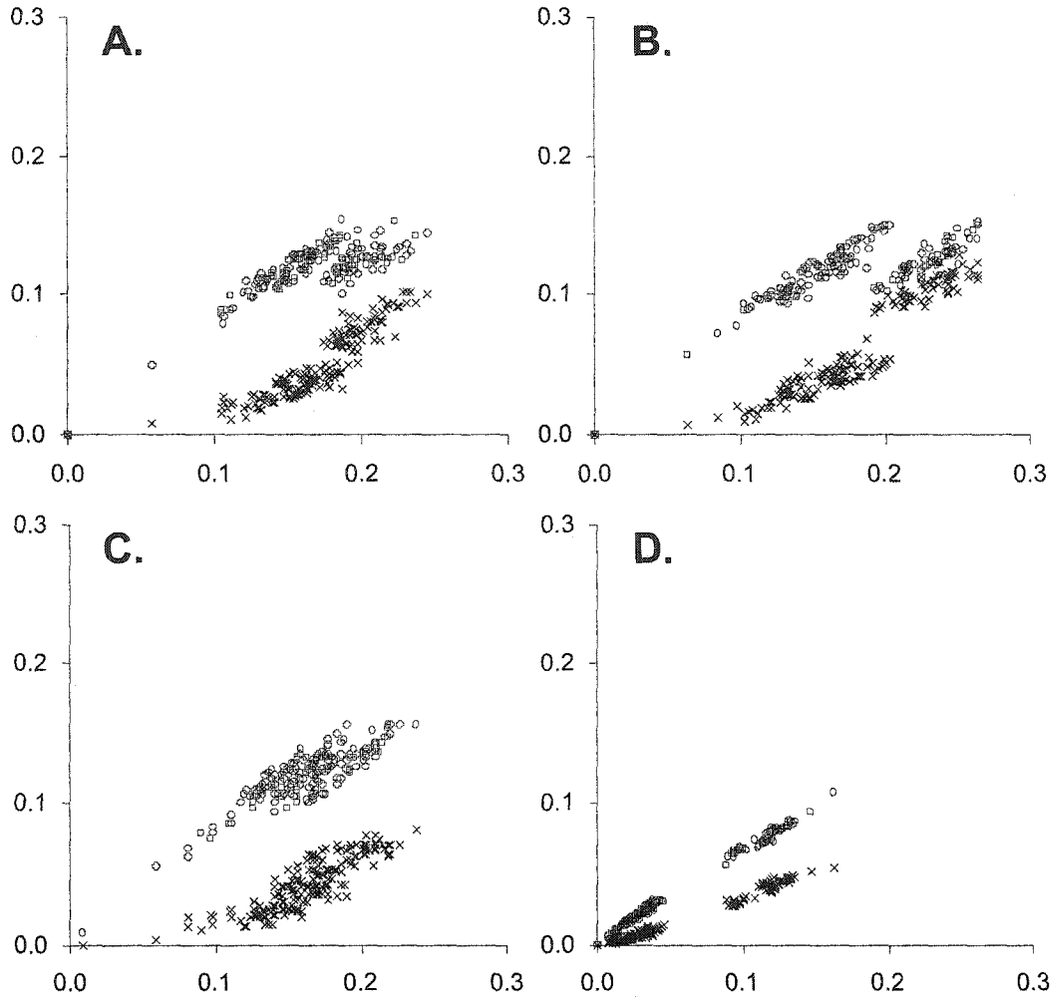


**Figure 4.3**

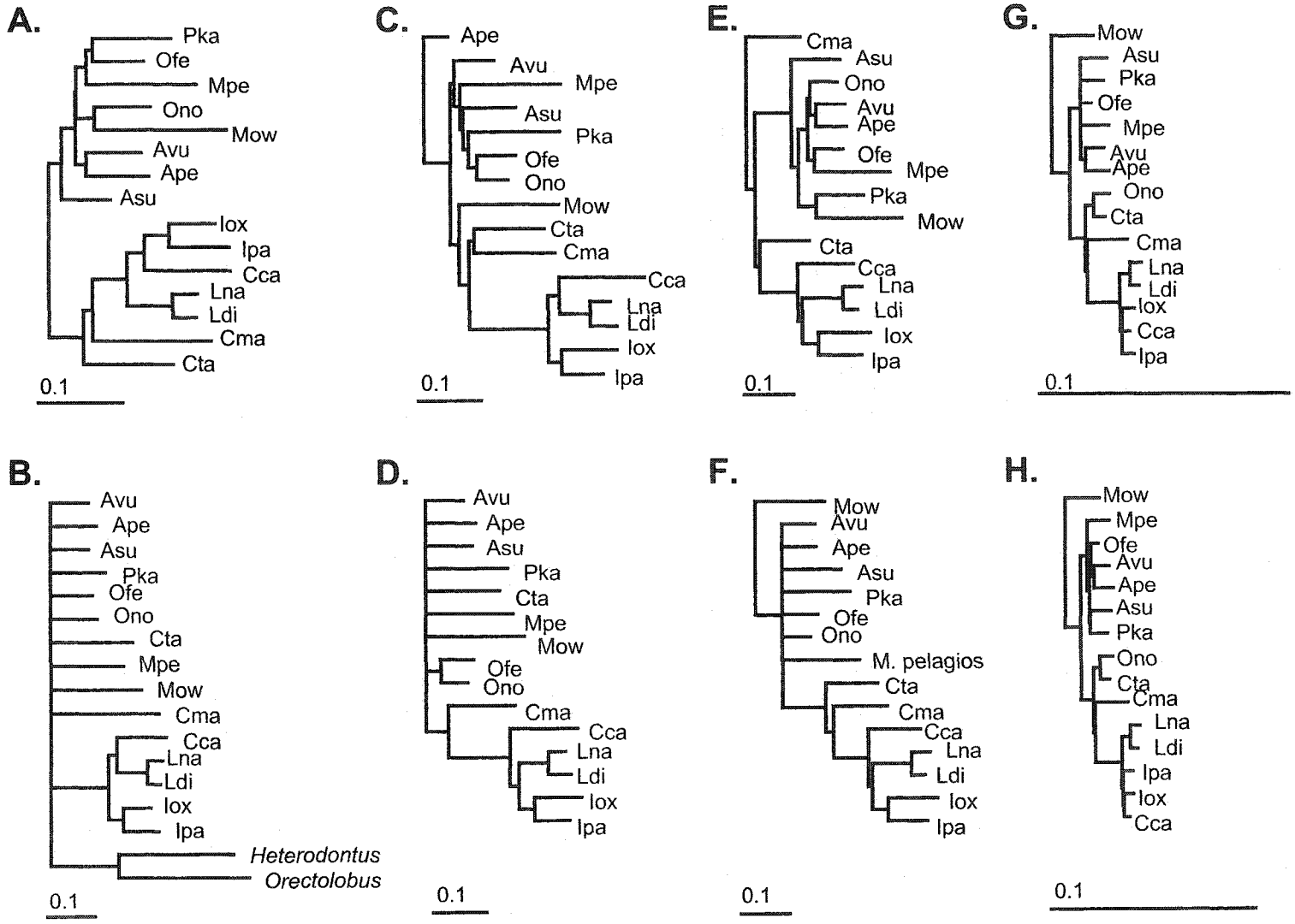
A.
Pka
Ofe
Mpe
Ono
Mow
Avu
Ape
Asu
Iox
Ipa
Cca
Lna
Ldi
Cma
Cta
0.1
C.
Ape
Avu
Mpe
Asu
Pka
Ofe
Ono
Mow
Cta
Cma
Cca
Lna
Ldi
Iox
Ipa
0.1
E.
Cma
Asu
Ono
Avu
Ape
Ofe
Mpe
Pka
Mow
Cta
Cca
Lna
Ldi
Iox
Ipa
0.1
G.
Mow
Asu
Pka
Ofe
Mpe
Avu
Ape
Ono
Cta
Cma
Lna
Ldi
Iox
Cca
Ipa
0.1
B.
Avu
Ape
Asu
Pka
Ofe
Ono
Cta
Mpe
Mow
Cma
Cca
Lna
Ldi
Iox
Ipa
Heterodontus
Orectolobus
0.1
D.
Avu
Ape
Asu
Pka
Cta
Mpe
Mow
Ofe
Ono
Cma
Cca
Lna
Ldi
Iox
Ipa
0.1
F.
Mow
Avu
Ape
Asu
Pka
Ofe
Ono
M. pelagios
Cta
Cma
Cca
Lna
Ldi
Iox
Ipa
0.1
H.
Mow
Mpe
Ofe
Avu
Ape
Asu
Pka
Ono
Cta
Cma
Lna
Ldi
Ipa
Iox
Cca
0.1

**Figure 4.4**

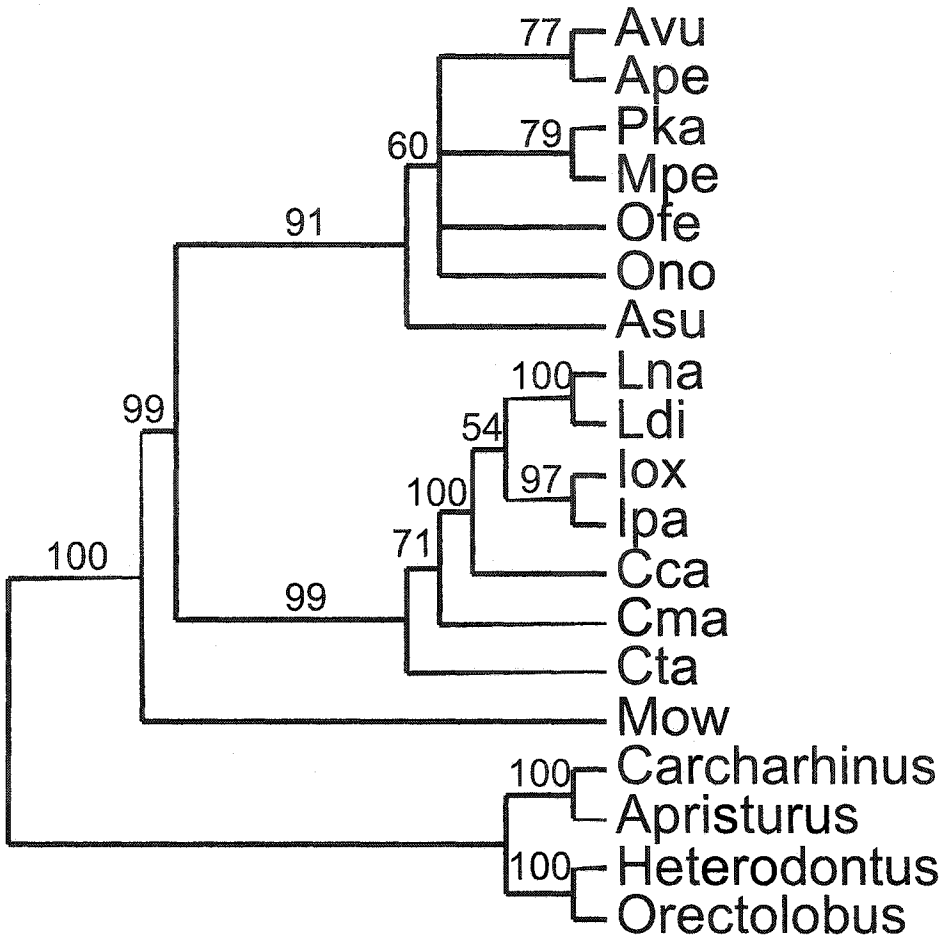
77
Avu
Ape
79
Pka
60
Mpe
91
Ofe
Ono
Asu
100
Lna
99
54
Ldi
97
Iox
100
Ipa
71
100
Cca
99
Cma
Cta
Mow
100
Carcharhinus
Apristurus
100
Heterodontus
Orectolobus

**Figure 4.5**

# CHAPTER 5. PHYLOGENETIC RELATIONSHIPS OF SHARKS OF THE FAMILY TRIAKIDAE (CARCHARHINIFORMES: ELASMOBRANCHII)

A paper to be submitted to *Copeia*

Juan Andrés López, Julie A. Ryburn, Olivier Fedrigo, Gavin J. P. Naylor

## Abstract

There are relatively few studies that examine the relationships between and within families of carcharhiniform sharks in a modern phylogenetic framework. The currently accepted classification schemes for these groups are largely based on the works of Compagno (1970, 1973), most recently updated and summarized in Compagno (1988a), whose conclusions combine phenetic and cladistic considerations. We have conducted a phylogenetic study of inter and intra-familial relationships of the shark family Triakidae (Carcharhiniformes) with the goal of testing the accepted hypotheses of relationships in this group. Our analyses and conclusions are based on information from the DNA sequences of four protein-coding genes (three from the mitochondrial genome and a single copy nuclear gene) from eight of the nine genera assigned to the Triakidae and 20 of the 39 species in the group. Of Compagno's proposed triakid clades, the sequence data offer strong support for the Galeorhinini (*Hypogaleus* and *Galeorhinus*); part of the Iagini (*Furgaleus* and *Hemitriakis* but not *Iago*); and part of the Triakinae (*Mustelus*, *Scylliogaleus* and part of *Triakis*). In addition, these data offer strong indication of paraphyly of the genera *Triakis* and *Mustelus*. The subgenera *Triakis* and *Cazon* of *Triakis* may not be sister taxa. *Mustelus* may not form a monophyletic genus unless *S. quecketti* and some species of *Triakis* are included. An expanded genus *Mustelus* includes two distinct clades diagnosed by mode of

reproduction. These sequences do not provide strong evidence for or against the monophyly of the family Triakidae.

**Introduction**

The family Triakidae (the houndsharks) is one of eight that form the most species-rich order of sharks, the Carcharhiniformes. As currently defined, the Triakidae includes extant representatives of an intermediate evolutionary phase between primitive carcharhiniforms (e.g. Scyliorhinidae – catsharks) and the derived and species-rich Carcharhinidae (Compagno, 1988b). The species assigned to the Triakidae are mostly small to medium in size. They are restricted to coastal regions in tropical and temperate seas throughout the world and feed primarily on benthic and mid-water crustaceans, cephalopods and bony fish. Many triakid species are targets of commercial fisheries of local to regional significance (e.g. Simpfendorfer et al., 2002). The relationships between species in the family are poorly understood due in part to the relatively scarcity of studies focusing on this question. To date, there is no strong body evidence supporting the monophyly of the family or even of the most species-rich genus in the family: *Mustelus*. In the present study, we use DNA sequences to examine existing hypotheses of phylogenetic relationships of triakid sharks.

White (1936) first gave family status to a group of carcharhinid taxa that included those currently assigned to the Triakidae. Prior to White's (1936) elevation of this group to family rank, its use in taxonomic classification can be traced back to Gray's (1851) tribe Triakiana of the family Squalidae (Compagno, 1984). White included in the Triakidae the carcharhinid genus *Triaenodon* (White, 1937), which created difficulties for subsequent workers attempting to delineate diagnostic characteristics for the family. As a temporary

solution, Compagno (1970) grouped triakids and carcharhinids in a single family. Later, Compagno (1973) resurrected and redefined the family Triakidae to include ten genera of carcharhiniform sharks grouped by a mosaic of characters, variably present and absent in the species ascribed to the group. Compagno (1973) included the following genera in the Triakidae: *Mustelus*, *Galeorhinus*, *Triakis*, *Scylliogaleus*, *Hemitriakis*, *Furgaleus*, *Hypogaleus*, *Iago*, *Gogolia* and the now invalid *Allomycter—A. dissutus* is considered a junior synonym of *M. canis* (Compagno, 1984; also discussed in Heemstra, 1973). More recently, Compagno (1988a) recognized the difficulty in defining diagnostic traits to characterize the Triakidae and discussed the possibility of its paraphyly. In the absence of alternative proposals, Compagno's (1988b) classification of carcharhiniform sharks families and genera is generally accepted.

Compagno (1988b) attempted to devise a classification of carcharhiniform taxa that reflected a cladistic interpretation of his morphological observations. This effort yielded ambiguous results concerning triakid relationships, thus leading Compagno to offer a provisional classification for triakid taxa and two alternative proposals of phylogeny (Compagno, 1988a). Currently, triakid classification divides the genera between two sub-families: the Triakinae and the Galeorhininae. The former includes the species in the genera *Triakis*, *Mustelus* and *Scylliogaleus*, and the latter, the species in the remaining genera of the Triakidae. Galeorhinin species were further classified into iagins and galeorhininins at the tribe level, with the species of *Iago*, *Furgaleus*, *Hemitriakis* and *Gogolia* included in the Iagini and the remaining galeorhinins in the Galeorhininini. The two alternative phylogenetic arrangements of triakid taxa offered by Compagno (1988a) differ in the monophyly of the family but conserve the sub-groupings outlined above (Fig. 5.1).

The lack of morphological evidence to unambiguously support the monophyly of the Triakidae prompted us to test Compagno's ideas of triakid relationships using evidence from DNA sequences. Here we report a phylogenetic investigation of triakid inter- and intra-family relationships based on analyses of DNA sequences from four protein-coding genes (3 mitochondrial and 1 single copy nuclear) from representatives of all but one of the nine genera and 20 of the 39 nominal species currently assigned to the family. The only triakid genus not included in this study is *Gogolia*, which is only known from type material preserved in formaldehyde.

## Materials and Methods

### *Specimen identification*

Tissue samples were obtained from freshly caught animals and preserved in 95% ethanol or a DMSO, Sodium Chloride, EDTA solution. In all cases, specimens were identified in the field using Compagno's (1984) and Heemstra's (1973, 1997) identification keys by G. Naylor or by other investigators experienced in the identification of the shark faunas of specific regions. In all cases two or more individuals from each species were included in our sample. In cases were species identification proved suspect; DNA sequences from putative conspecifics identified by different individuals and collected from different localities were compared to determine whether or not the nominal species formed a readily diagnosed taxonomic unit. In addition, five specimens from undescribed or undeterminable species of triakids were included among the specimens examined to determine their affinities to known triakid species. These five undetermined taxa were: *Hemitriakis* sp. identified as a previously undescribed species from Japan by K. Nakaya (pers. comm.); *Iago* sp. from the Philippines, distinguishable from described species in the genus; and, three distinct groups of

specimens of *Mustelus* from Borneo, Baja California and the Gulf of Mexico. Because the monophyly of the Triakidae is in question, we included potential outgroups from primitive and advanced carcharhiniform families. The triakid and outgroup taxa included in this study are listed in Table 5.1. Despite repeated attempts, we were unable to obtain DNA from the formalin-preserved specimens that constitute the only known record of the triakid genus *Gogolia.*

*Amplification and Sequencing*

We obtained complete coding sequences for the mitochondrial genes Cytochrome b (Cyt *b*) and NADH Dehydrogenase subunit 2 (ND2); and partial coding sequences for NADH Dehydrogenase subunit 4 (ND4) and the single copy nuclear gene *recombination-activating gene* 1 (RAG1). In each case, we used amplification and sequencing primers specifically designed to target a wide diversity of species of galeomorph sharks. New primers were designed as needed when amplification or sequencing of specific genes and taxa proved problematic. PCR amplification conditions were optimized for each primer set and, in some cases for problematic taxa. In general, annealing temperatures ranged between 48C and 55C, and the number of cycles ranged between 32 and 40. Detailed amplification protocols for specific taxa and gene are available from the authors upon request.

We purified PCR products by centrifugation through size selective filters (Millipore, MA) according to manufacturer's recommendations. We then used the purified PCR products as templates in chain-termination reactions using fluorescently labeled chain terminators (PerkinElmer, MA). We purified the chain-termination reaction products by ethanol precipitation and submitted this purified products to the DNA Synthesis and Sequencing facility at Iowa State University for sequencing determination by electrophoresis

on automated sequencing machines (PerkinElmer, MA). Table 1 lists the species and gene sequences included in the analyses.

*Analyses*

We aligned the sequences with the aid of the computer program Clustal X (Thompson et al., 1997) with default settings. Then, we manually edited the computer generated alignments using the amino acid translation of the DNA sequences to guide the placement of gaps to preserve the reading frame and to correct obvious artifacts of the automated alignments. We excluded all missing or ambiguous characters from all subsequent analysis. Prior to performing phylogenetic analysis, we tested the hypothesis of base composition homogeneity for each gene using the $\chi^2$-test implemented in PAUP* (Swofford, 1998) to determine if deviations from stationarity in this parameter had the potential to influence the results of phylogenetic analyses. We also plotted the proportion of transition and transversion substitutions against total uncorrected pairwise sequence divergence to determine if the divergence between the sequences examined had reached saturation levels and therefore may be expected to be affected by the deterioration the historical information present in the sequences at clade origin.

For each set of gene sequences, we obtained phylogenetic trees based on Maximum Parsimony (MP), distance (D), and Maximum Likelihood (ML) criteria of optimality using the computer program PAUP* (Swofford, 1998). To determine the optimal tree under MP, we use the following heuristic search settings: 1000 replicates with random taxon addition and random starting trees with tree-bisection-reconnection (TBR) branch swapping and saving the optimal tree from each replicate. Under MP and D optimality we also obtained the consensus tree of the best trees from 1000 bootstrap pseudo-replicates of the data sets. In

addition, we conducted the same series of analyses on a data set that combined the four gene sequences for those species from which all four sequences were obtained. To arrive at a proposal for triakid relationships, we examined the tree topologies resulting from all the different analysis in search for clades that were consistently and robustly supported by all data sets and method of phylogenetic reconstruction. Aside from the consistent and robust clades supported in this study, we defer to Compagno's (1988a) proposals for triakid classification.

## Results

### *Cytochrome b*

We determined the complete coding sequence of Cyt *b* for the triakid species listed in Table 1. Excluding missing and ambiguous sites, the Cyt *b* alignment includes 857 sites. Base composition homogeneity is rejected only when third codons position sites are examined independently ($p < 0.01$), but even at these sites, homogeneity is not rejected when the outgroup species are excluded from the test (Table 5.2). As expected, plotting substitutions against total divergence at all sites and by codon position shows that the substitution characteristics of the third codon positions differ markedly from those of the rest of the sequence (Fig 5.2). These plots do not reveal extensive substitution saturation between triakid taxa; however some of the points in the plots based on comparisons involving outgroup species indicate changes in the substitution process at third codon position sites. The fact that these changes seem to be restricted to third codon position sites and to comparisons with outgroup taxa suggest that they are related to the observed deviation from base composition homogeneity.

Parsimony analyses of the uniformly weighted Cyt *b* sequences show strong support for several of the clades of triakid taxa that were proposed by Compagno (1988). The sister group relationship between *Galeorhinus* and *Hypogaleus* is strongly supported in the bootstrap analysis (99%). Also well supported is the sister group relationship between *Furgaleus* and *Hemitriakis* (87%), which are two of the four genera of the Iagini. The other iagin genus included in this study, *Iago*, is not supported as a close relative of the *Furgaleus* + *Hemitriakis* clade. *Scylliogaleus quecketti*, *Triakis megalopterus* and all the species of *Mustelus* examined form a strongly supported clade (90%) in which the genus *Mustelus* is paraphyletic. Among species of *Mustelus* included in this study, the Cyt *b* sequences offer strong support (93%) for a clade including *M. canis*, *M. californicus*, *M. mosis*, *M. mustelus* and *M. norrisi*. The other species of *Mustelus* (*M. manazo*, *M. asterias*, and *M. schmitti*) form a strongly supported clade (100%) that forms and unresolved polytomy with *S. quecketti* and *T. megalopterus*. The other two species of *Triakis* in this study (*T. scyllium* and *T. semifasciata*) are supported as sister species in this analysis (99%) and are never placed close to the clade that includes their congener *T. megalopterus* (Fig. 5.3).

Aside from the groups described above and some groups of species of *Mustelus* and of outgroup taxa, the bootstrap consensus tree offers no resolution for nodes deep in the tree. The monophyly of the Triakidae is neither strongly supported nor strongly contradicted by Cyt *b* sequences; although it is not found in any of the optimal trees obtained from these sequences. The two alternative topologies proposed by Compagno (1988; Fig. 5.1) show insignificant differences in number of steps and likelihood scores. The relationships of the triakid genera to more derived and more primitive carcharhiniforms are satisfactorily resolved (i.e. consistent and robust support) in any of the phylogenetic analyses of cytochrome *b* sequences. Further, the hypothesis supported by uniformly weighted

parsimony analysis of these data does not support the monophyly of the carcharhinid outgroups, although an analysis where only transversions are considered at third codon position sites offers support for the monophyly of the carcharhinid outgroups minus *Galeocerdo cuvier*, which is in agreement with results of an ongoing study of carcharhinid phylogeny (G. Naylor, unpublished data).

In the process of collecting the triakid Cyt *b* sequences, we discovered a partial gene duplication of a 3' region of the Cyt b gene located downstream of the functional gene. We have established the presence of the duplication only in *Hemitriakis japanica* and *Mustelus asterias*; however the methods we employed do not allow us to rule out its presence in other triakids. The Genbank accessioned sequence of the complete mitochondrial genome of *M. manazo* (NC000890; Cao et al., 1998) shows that this close relative of *M. asterias* lacks the duplication. Sequence analysis of the duplicated region shows that in both *H. japanica* and *M. asterias* the redundant sequence originated from the same part of cytochrome *b* with small differences at the duplication boundaries. A phylogenetic analysis of the functional and duplicated Cyt *b* sequences strongly supports the hypothesis of independent origins of the duplication in *H. japanica* and *M. asterias*. Thus these duplicated regions represent a striking case of the parallel emergence of a rare mitochondrial genome rearrangement and suggest a common mechanism of origin.

*NADH2*

We determined the complete coding sequence of ND2 for the triakid species listed in Table 5.1. Excluding missing and ambiguous sites, the ND2 alignment includes 1244 sites. The results of chi-square tests of base composition homogeneity are similar to those obtained for the Cyt *b* sequences. Over all sites the tests fail to reject stationarity, but when the third

codon position sites of triakid and outgroups taxa are considered separately, the test rejects the hypothesis of homogeneity. This effect is not observed when outgroup taxa are excluded from the test (Table 5.2). The plots of transitions and transversions against uncorrected distance present an interesting contrast to those based on the Cyt *b* sequences (Fig 5.4). Like those, when all sites are considered, there is little evidence of substitution saturation except for some comparisons involving outgroup taxa. However, in the ND2 sequences the evidence of substitution heterogeneity in outgroup comparisons is most clearly observed at first and second codon position sites and is restricted to comparisons involving the outgroups *Poroderma pantherinum* and *Leptocharias smithii*. The substitution plots of ND2 and Cyt *b* sequences show highly contrasting patterns of dispersion, especially at first and second codon position sites, where the Cyt b sequences are broadly dispersed at all levels of divergence while the ND2 sequences follow a homogeneous substitution pattern (Figs. 5.2B, C and 5.4B, C). The shape of the ND2 substitution plot for third codon position sites suggests that most divergent sequences in the sample have reached substitution saturation levels.

The results of phylogenetic analyses of ND2 sequences are in general agreement with those from the Cyt *b*. These data offer strong support for the three clades that group: (1) *Furgaleus* and *Hemitriakis*, (2) *Galeorhinus* and *Hypogaleus*, and (3) *Triakis megalopterus*, *Scylliogaleus quecketti*, and the species of *Mustelus*. Among species of *Mustelus*, there is strong support for two clades: the 'mustelus' clade, which includes *M. californicus*, M. *canis*, *M. henlei*, *M. mosis*, *M. mustelus*, and *M. norrisi*; and the 'asterias' clade, which groups *M. antarcticus*, *M. asterias*, *M. manazo*, *M. schmitti*. The monophyly of *Mustelus* is contradicted by the inclusion of *T. megalopterus* and *S. quecketti*, which, together as a clade,

are placed as the sister group of the 'asterias' clade of species of *Mustelus*, with weak support (Fig. 5.5).

The tree topologies supported by parsimony analysis of the ND2 data do not provide strongly supported resolution of the relationships between the triakid genera and the outgroups. These data offer moderate support for a clade formed by *Leptocharias*, the triakids and the higher carcharhinoids (75%), but offers no resolution for relationships between those lineages. There are only small differences in numbers of steps and likelihood scores between the two alternative topologies proposed by Compagno (1988; Fig. 5.1). However, in distance-based bootstrap analysis of these sequences the triakid taxa form a monophyletic clade in 68% of the replicates. That analysis also offers weak support for the monophyly of the advanced carcharhiniform outgroups (52%).

*NADH-4*

We determined the partial coding sequence of ND4 and its 3' flanking region for the triakid species listed in Table 5.1. The ND4 sequence alignment includes 831 sites, which encompass 645 sites at the 3' end of the ND4 gene and the sequences of the histidine and serine transfer RNA genes located downstream of the ND4 gene. There is no evidence of base composition bias in the ND4 sequences even when third codon position sites are considered separately (Table 5.2). The substitution plots do not indicate extensive substitution saturation in comparisons between triakid species (Fig. 5.6). Like the substitution plots of the two other mitochondrial genes, the ND4 plots show a distinct pattern of substitution for the comparisons involving the most divergent outgroups. Otherwise, the plot shows continually increasing levels of both types of substitutions with limited point dispersion, which suggest a homogeneous substitution process.

The results of phylogenetic analyses based on ND4 sequences (Fig. 5.7) are in general agreement with those supported by the analyses based on Cyt *b* and ND2. Among the clades that find consistent support in these data are the Galeorhinini (92%), a clade composed of *Furgaleus* and *Hemitriakis* (68%), a clade composed of *Triakis megalopterus* and *Scylliogaleus quecketti* (92%), and a clade that includes these last two species and all of the species of *Mustelus* in our sample (97%). Also, in accordance with the results based on the other sequences, the species of *Mustelus* fall into two well-defined and strongly supported clades. The 'asterias' clade contains the species: *M. antarcticus*, *M. asterias*, *M. manazo* and *M. schmitti*. The 'mustelus' clade includes: *M. canis*, *M. henlei*, *M. mosis*, *M. mustelus*, *M. norrisi*, and *M. sinusmexicanus*. Relationships within these clades of species of *Mustelus* remain largely unresolved.

There is no significant support for the monophyly of the Triakidae in any of the analysis. Both parsimony and distance bootstrap trees consist of a largely unresolved polytomy that includes several consistently supported clades. In agreement with some of Compagno's ideas on the affinities of triakid taxa, the maximum likelihood analysis of ND4 sequences results in the placement of the Galeorhinini as the sister group of the carcharhinid taxa in our sample; however, the hemigalids are not part of this crown group as is proposed in Compagno's classification. Further, this placement of the galeorhininins is not found in other analyses of these sequences or in the results from other sequences.

*RAG-1*

We obtained partial sequences of the RAG1 gene for the species in Table 5.1. To maximize the number of taxa represented in the RAG1 alignment, we had to eliminate many regions of the gene that presented sequencing challenges. Excluding these missing and

ambiguous regions, the alignment contains 1015 sites of RAG1 coding sequence. The amount of divergence of the RAG1 sequences between triakid species is very low. The greatest observed differences between triakid taxa amounted to only slightly more than 3% when all sites are considered and 7% when third codon position sites are examined separately. As expected from the very low levels of divergence there are no detectable changes in base composition among the taxa in this study (Table 5.2). The substitution plots of the RAG1 sequences show a complex pattern of substitution characteristics (Fig. 5.8). The heterogeneity can be traced to third codon position sites (Fig. 5.8B), where the substitution characteristics between triakids are different than those between triakids and the outgroups. Further, the outgroup taxa show at leas two distinct substitution patterns. Although, these plots show no evidence of substitution saturation, they indicate shifts in the substitution process underlying the evolution of these sequences. Depending on the method of analysis, these shifts may affect the reliability of the phylogenetic inferences supported by these data.

Phylogenetic analysis of the RAG1 sequences result in unresolved relationships for the majority of the taxa in our sample (Fig 5.9). The lack of resolution between triakid taxa is not unexpected given the very low levels of divergence observed in those sequences. In the bootstrap analysis, strong support is restricted to: the galeorhininins (88%), a clade composed of all of the species of *Mustelus* in our sample, *Triakis megalopterus* and *Scylliogaleus quecketti* (95%), and, within this clade, a group formed by species of *Mustelus* from the 'asterias' clade (94%). The species of *Mustelus* from the 'mustelus' clade are grouped in a clade that receives marginal bootstrap support (61%). The clade formed by *Furgaleus* and *Hemitriakis* is not part of the bootstrap consensus tree.

In agreement with the results based on the mitochondrial sequences, the monophyly of the Triakidae is not found in any of the optimal trees or in the bootstrap consensus trees. However, Compagno's (1988) alternative proposals of triakid phylogeny do not differ in number steps and differ only slightly in likelihood scores. In the bootstrap consensus tree and in optimal trees, triakid monophyly is broken by the inclusion of *Leptocharias smithii.*, while the 'higher carcharhinoids' are supported as a monophyletic group (83%). This is in contrast with results based on the mitochondrial sequences where 'higher carcharhinoids' were found paraphyletic. Perhaps significantly, the branch length leading to *Leptocharias* is significantly longer than that of all other taxa in the sample with the exception of that of *Poroderma.* These outgroup taxa may be the source of the substitution heterogeneity observed in the substitution plots, thus making the accuracy of their placement on the phylogenetic reconstruction difficult to gauge. *Poroderma* does not group with the other scyliorhinid species in the sample.

*The 'unassigned' specimens*

Specimens collected from Philippines were identified as members of the genus *Iago* but could not be assigned with confidence to either of the two recognized species in the genus. We obtained mitochondrial DNA sequences from some of these specimens and compared them to those of specimens of *Iago omanensis.* The Cyt *b* sequences of the specimens of *Iago sp.* cluster with *I. omanensis* but the sequences from the two species are clearly distinct. They differ in 10.9% of the sites (9.1% transitions and 1.8% transversions). The ND2 sequences from *Iago* sp. are strongly supported as the sister lineage of those from *I. omanensis.* On average, the ND2 sequences from these two groups differ at 9.6% of the sites compared. Similarly, the ND4 sequences of the Philippine samples support the results

obtained from the other sequences: the Philippine sequences invariably group with sequences from *I. omanensis* but are 9.1% divergent from those.

Specimens collected in the vicinity of Ryukyu Island, Japan and examined by K. Nakaya were determined to belong to the genus *Hemitriakis* but also to be morphologically distinct from any of the four described species of that genus. The ND4 sequences from the specimens of *Hemitriakis* sp. group with those of *H. japanica*, but differ from them at 4-5% of the sites. It is interesting to note that the observed differences between these sequences reveal a very high ratio of transitions to transversions (range from 28-35).

Specimens of a species of *Mustelus* collected by Scott and Rachel Cavanaugh were assigned to *Mustelus* species A in reference to the taxa reported by Last and Stevens (1994) under that name from the shark fauna of Australia. There is an interesting pattern of divergence between the Cyt *b* sequences of these specimens and other species of *Mustelus*. The sequences of the Borneo specimens are approximately equidistant to those of *M. mosis* (2.56%) and *M. Mustelus* (2.73%), while the divergence between the sequences of last two species is 4.26%. However, geographically the Borneo specimens are closer to the range of distribution of *M. mosis* (western Indian, Red Sea and to Natal, South Africa; Compagno, 1984) than to that of *M. mustelus* (eastern Atlantic and Mediterranean: Compagno, 1984). The ND2 sequences from the Borneo specimens are closer to those from *M. mosis* (1.9%) than to those of *M. mustelus* (3.4%). Similarly, the ND4 sequences of the Borneo specimens are most similar to those from *M. mosis* (2.3%) and to those from *M. Mustelus* (3.2%). In phylogenetic analyses, sequences from the Borneo specimens cluster with those from *M. mosis* and, less consistently, with those from *M. mustelus*.

Two specimens of *Mustelus* that were collected from the Gulf of California by J. Caira were identified as *M. intermedius* and *M. platyrhinus* (Heemstra, 1973) based on the

number of precaudal vertebrae and the position of the onset of monospondyly in the vertebral column. The Cyt *b* and ND2 sequences from those two specimens differ in only 0.25% of the sites and form a highly supported group in all phylogenetic analyses. The group formed by these two sequences is nested within the 'mustelus' clade of *Mustelus*, but its relationships to other taxa in this group are unresolved.

Two other specimens of *Mustelus* collected from the Gulf of California were difficult to identify based on morphological traits. The mitochondrial DNA sequences of these specimens are most similar to those of *M. canis* (2.4 – 2.8%) and to those of *M. henlei* (3.0 – 3.8%). In all bootstrap analyses, there is strong support for a group composed of the specimens from the Gulf of California and *M. canis*, which is restricted to the Western Atlantic and Caribbean. This clade forms part of the 'mustelus' clade of *Mustelus*.

**Discussion**

Although several species of triakid sharks are commercially important, and as a result have been the focus of studies on different aspects of their biology (e.g. Donley and Shadwick, 2003; Conrath et al., 2002; Hamlet et al., 2002); there has been relatively little interest in the phylogenetic relationships of this widespread family of sharks. The present study is the first comprehensive attempt to test ideas of triakid inter- and intra-relationships since the group was redefined and examined by Compagno (1973; 1988a). A close examination of Compagno's (1988a) work on triakid relationships reveals that his proposal for triakid systematics does not derive objectively from a character state data matrix, but rather appears to be the result of a thorough understanding of carcharhiniform anatomy that is not presented in a explicit phylogenetic argument based on a character state matrix. This makes the robustness of Compagno's results difficult to gauge. In this respect, the sequence-

based phylogeny represents an improvement over the currently held views of triakid classification. On the other hand, the sequence-based phylogeny contains many unresolved relationships; therefore there are aspects of triakid classification that may remain provisionally arranged according to Compagno's (1988a) hypotheses. The results of phylogenetic analyses based on DNA sequences agree in several respects with Compagno's arrangement of triakid taxa but differ from those in some critical aspects that have implications for the taxonomy of triakid taxa. The agreements and differences between the morphology and sequence-based hypotheses are highlighted below.

*Systematic implications for* Mustelus, Triakis *and* Scylliogaleus

Our sample included three species of the genus *Triakis* (*megalopterus*, *scyllium* and *semifasciata*). None of our results support a monophyletic *Triakis*. On the contrary, all of our analyses consistently place *Triakis megalopterus* and *Scylliogaleus quecketti* as sister species. Further, our results indicate that this clade is most closely related to a well-defined group of species of *Mustelus*, the 'asterias' clade. This arrangement is one of the most highly and consistently supported non-trivial clades found in our results. These results suggest that *Triakis* and *Mustelus* as currently delineated constitute paraphyletic assemblages and that there may be valid practical considerations to eliminate the monotypic genus *Scylliogaleus*.

The inclusion of *T. megalopterus* in the clade formed by the species of *Mustelus* is not completely unexpected. The five species of the genus *Triakis* have been divided into two subgenera *Triakis* (for *T. semifasciata* and *T. scyllium*) and *Cazon* (for *T. acutipinna*, *T. maculata*, and *T. megalopterus*). If our results are correct, the two subgenera may not form a monophyletic group and the species of *Cazon* may be more appropriately assigned to *Mustelus*; pending evidence supporting the monophyly of *Cazon*, which cannot be derived

from this study due to incomplete taxon sampling. Compagno (1988a) posited an affinity between *Cazon* and *Mustelus* based on features of the cranium and dentition. Similarly, the placement of *S. quecketti* in our results suggests that the species should be reassigned to the genus *Mustelus*, where it groups with *T. megalopterus* and the 'asterias' clade of *Mustelus*. Compagno (1988a) proposed a close relationship of *Scylliogaleus* to *Mustelus* and *Cazon*. He also noted that among the species of *Cazon*, *T. acutipinna* shows expanded sphenopterotic ridges, which is a trait also observed in *S. quecketti*. This shared trait may be evidence of a sister species relationship between the two species, or of the close affinity between *Scylliogaleus* and *Triakis* supported by the sequence data.

In summary, the placement of *T. megalopterus* and *S. quecketti* in the phylogenetic hypothesis supported by DNA sequences finds some support in morphological observations of Compagno (1988a), although it contradicts the conclusions reached in that study. If the sequence-based placement of these taxa is accepted then the genus *Triakis* would require redefinition. Because the type species of the genus (*T. scyllium*) is not nested in the 'asterias' clade of *Mustelus*, the redefined genus may conserve the name without creating taxonomic confusion. The nomenclatural treatment of the displaced species of *Triakis* would depend on the revised taxonomy of *Mustelus*, as discussed below. Similarly, the fate of the genus *Scylliogaleus* would depend on the redefinition of the genus *Mustelus*.

*The genus* Mustelus

Among triakid genera, *Mustelus* is not only the most species-rich but also the most problematic with respect to taxonomy and systematics. The most recent and comprehensive attempts to clarify taxonomic issues in the genus show that in many cases species differences are difficult to delineate (Heemstra, 1973; 1997) due to the scarcity of specimens available

for study and, perhaps more significantly, to the minimal divergence evident in their morphologies. Our study is aimed at improving our understanding of triakid inter-generic relationships so the systematics of the genus *Mustelus* fall outside of its scope. However, the sample of species of *Mustelus* that represent that genus in our study allows us to draw certain conclusions on the taxonomy of *Mustelus*.

The DNA sequences we analyzed in this study offer strong support for the monophyly of an expanded *Mustelus* (with the inclusion of *T. megalopterus* and *S. quecketti*) and for the subdivision of the expanded genus into two groups. In our taxonomic sample, this categorization delineates a group of species with viviparous reproduction with a yolk-sac placenta (the 'mustelus' clade) and one with ovoviviparous reproduction (the 'asterias' clade). Because our sampling of species of *Mustelus* is incomplete, we cannot yet determine if this trait will clearly define two sub-genera within the group. If this division is supported by future studies that include a greater diversity of species of *Mustelus*, then the significance of the different mode of reproduction between the two groups of *Mustelus* may justify dividing the genus in two. Under such a scenario, members of the 'mustelus' clade would conserve their current generic designation because the type species (*M. mustelus*) is a member of that clade. The members of the 'asterias' clade, including *T. megalopterus* and *S. quecketti* may be assigned to a new genus. From a taxonomic perspective, a more conservative alternative would be to reassign *Scylliogaleus* and the species of *Cazon* (assuming this subgenus is monophyletic) to an expanded *Mustelus*.

Aside from the two well-defined clades, there are only a few consistently supported relationships among the species of *Mustelus* in our study. One of these is the repeated grouping of *M. mustelus*, *M. mosis* and unassigned specimens of *Mustelus* collected in Borneo. This group of taxa includes all of the members of the 'mustelus' clade that do not

occur on the coasts of the American continent. However, the relationship of this group to other taxa in the 'mustelus' clade remains unresolved. Another group consistently found within the 'mustelus' clade is formed by specimens of *M. canis* from the Atlantic coast of the U.S. and unassigned specimens of *Mustelus* collected from the Pacific coast of Mexico, where *M. canis* does not occur. Overall, the magnitude of divergence between the sequences from species of *Mustelus* is relatively low, which is in agreement with the relatively minor morphological characteristics that distinguish many of the species in this genus. Unfortunately, the low levels of divergence among species of *Mustelus* indicates that a full understanding of the taxonomy and systematics of the genus will require extensive knowledge of the genetic diversity of populations and species to be able to justify the definition of species boundaries. Gardner and Ward (2002) used this approach to better understand the genetic diversity some of the species of *Mustelus* from Australia, however the taxonomic sample of the study is poor, which limits the generality of the conclusions they reach. In addition, given the low levels of sequence divergence and the overlap in geographic distribution of many species of *Mustelus*, it is likely that hybridization and incomplete lineage sorting occurs in the genus, thus increasing the complexity of the task of delineating species and devising methods for species identification in the field. Under these circumstances, thorough taxonomic samples and big sample sizes for each taxonomic unit are necessary.

*Triakid monophyly*

As a result there are no diagnostic or synapomorphic characters that group the taxa assigned to the Triakidae. Consequently, Compagno offered two arrangements of the triakid taxa. In one of these proposed arrangements the triakids form a paraphyletic group of three

lineages diverging from the carcharhiniform stem after the origin of the Leptochariidae and before the splitting of the hemigalid and carcharhinid lines. Alternatively, the monophyletic Triakidae is the sister group of the 'advanced' carcharhiniforms (hemigalids and carcharhinids).

Our analyses produced equivocal evidence regarding triakid monophyly. Overall our phylogenetic analyses do not support a monophyletic Triakidae, however there is no alternative arrangement that is consistently supported by the different data sets. Interestingly, although the low levels of divergence that we observed in our sequences and the homogeneous substitution patterns evident in the substitution plots led us to expect well resolved tree topologies, all four genes are ambiguously informative about the deeper levels of triakid relationships. The results of this study are silent regarding triakid monophyly; therefore Compagno's (1988a) views on this subject remain untested and may be provisionally accepted in the interest of taxonomic stability.

*The unassigned specimens*

The DNA sequence evidence shows that the unidentified specimens of *Hemitriakis* collected around the Ryukyu Island of Japan in the same region inhabited by *H. japanica* represent a distinct taxon not represented by any other specimen in our study. The specimens of *Hemitriakis* sp. were collected and examined by Dr. K. Nakaya, who determined that they did not fit the diagnostic characteristics of any known species of *Hemitriakis* (summarized in Compagno and Stevens, 1993). The specimens from Japan can be distinguished from the Australian species of *Hemitriakis* by differences in vertebral counts, morphometrics and color pattern in the young. Also, it may be significant that the specimens from Japan and the Australian species of *Hemitriakis* are each parasitized by different species of copepods of the

genus *Lernaeopoda* (M. Takahashi, pers. comm.). We now have corroborative evidence from DNA sequences that these specimens represent a genetically distinct entity, which may bring the total number of species of *Hemitriakis* to five. The taxonomic sample in the present study prevents us from making statements about relationships among the species of *Hemitriakis*.

The specimens of *Iago* from the Philippines also seem to belong to a distinct taxon related to *Iago omanensis*. The collectors specifically stated that these specimens do not correspond to *I. garricki*, which was described from Vanuatu and may extend to northwestern Australia (Fourmanoir and Rivaton, 1979; Compagno, 1984). The mitochondrial sequences of these two taxa show considerable differences (9%) but they are always supported as sister species. Compagno (1984) noted that specimens of *Iago* from this region, although similar to *I. garricki* probably corresponded to an undescribed species. Our species sample did not include specimens of *I. garricki* therefore we cannot determine with certainty the relationships between that species and the specimens from the Philippines, however together with the distinct morphological characteristics of these individuals, the sequence data supports Compagno's (1984) report of an undescribed species from the area.

The phylogenetic placement of the specimens from Borneo and the low level of sequence divergence between them and the allopatric *M. mustelus* and *M. mosis* highlight some of the problems associated with understanding and defining biological species in this genus. Because the geographic distributions of these three taxa do not overlap and their morphologies are so similar, it is necessary to have a more thorough sample of the genetic diversity in each taxon before defining the geographic, genetic and morphological boundaries of each species. Similarly, specimens of *Mustelus* from the Gulf of California that were identified as *M. intermedius* and *M. platyrhinus* have almost identical mitochondrial DNA

sequences, which highlights the possibility of hybridization among species of *Mustelus*. Due to the uncertainties associated with identifying closely related species of *Mustelus* we refrain from offering hypothesis regarding the species affinities of the unidentified specimens that were assigned to that genus.

*Conclusions*

In agreement with the conclusions reached by Compagno (1988), the DNA sequences offer ambiguous evidence regarding the monophyly of the Triakidae. We can only state with confidence that if the genera currently assigned to the group indeed are members of a natural clade, then the evidence of that shared ancestry is exceedingly scarce both in their anatomy and in the DNA sequences of four of their genes. Based on this study we propose changes to the current hypothesis of triakid relationships based only on the strongly supported aspects of our results. Specifically, the subgenus *Cazon* of *Triakis* belongs in the 'mustelus' clade of species of *Mustelus*. Provisionally, the single species of the genus *Scylliogaleus* is the sister species of the species of *Cazon* but Compagno (1988) reported evidence that implies a paraphyletic *Cazon* with respect to *Scylliogaleus*. The species of *Mustelus* belong to two distinct clades that may be defined by the mode of reproduction. *Furgaleus* and *Hemitriakis* are sister genera but our results do not support the inclusion of these two in the Iagini. As proposed by Compagno (1988), *Galeorhinus* and *Hypogaleus* form a well-supported monophyletic clade. Like Compagno (1988) we remained uncommitted regarding the monophyly of the Triakidae.

We examined four different genes to minimize the effects that the violation of assumptions implicit in phylogenetic reconstruction schemes by any given gene could have on our conclusions. All methods of phylogenetic reconstruction fail to infer the true pattern

of evolutionary relationships among terminal constructs, be they assemblages of nucleotides, amino acids or morphological traits, when certain conditions are not met by those entities (e.g. Felsenstein, 1978; Lockhart et al., 1994). These requisite conditions vary by method of inference. By comparing results from different analyses of the different genes and gene combinations and deriving our conclusions from a conservative interpretation of those comparisons we hope to exclude spurious inferences from our conclusions on triakid relationships. Given all the known pitfalls associated with phylogenetic reconstruction we prefer proposing a conservatively interpreted hypothesis of relationships even if it contains numerous unresolved polytomies.

**Acknowledgements**

The work presented here was possible thanks an NSF grant to GJPN. Also, we are indebted to colleagues throughout the world who generously provided samples of species of triakids. J. Caira and her reseach group contributed an important number of samples for this study.

**Literature Cited**

CAO, Y., P. J. WADDELL, N. OKADA, AND M. HASEGAWA. 1998. The complete mitochondrial DNA sequence of the shark *Mustelus manazo*: Evaluating rooting contradictions to living bony vertebrates. Mol. Biol. Evol. 15:1637-1646.

COMPAGNO, L. J. V. 1970. Systematics of the genus *Hemitriakis* (Selachii: Carcharhinidae), and related genera. Proceedings of the California Academy of Sciences. 38:63-98.

—. 1973. Gogolia filewoodi, a new genus and species of shark from New Guinea (Carcharhiniformes; Triakidae), with a redefinition of the family Triakidae and a key to the genera. Proceedings of the California Academy of Sciences. 39:383-410.

—. 1984. FAO species catalogue. Vol. 4, Sharks of the world. An annotated and illustrated catalogue of shark species known to date. FAO.

—. 1988a. Family Triakidae Gray, 1851: Houndsharks, p. 206-252. In: Sharks of the order Carcharhiniformes. Princeton University Press, Princeton, New Jersey.

—. 1988b. Sharks of the order Carcharhiniformes. Princeton University Press, Princeton, New Jersey.

COMPAGNO, L. J. V., AND J. D. STEVENS. 1993. *Hemitriakis falcata* n. sp. and H. *abdita* n. sp., two new Houndsharks (Carcharhiniformes: Triakidae) from Australia. Records of the Australian Museum. 45:195-220.

CONRATH, C. L., J. GELSLEICHTER, AND J. A. MUSICK. 2002. Age and growth of the smooth dogfish (*Mustelus canis*) in the northwest Atlantic Ocean. Fishery Bulletin. 100:674-682.

DONLEY, J. M., AND R. E. SHADWICK. 2003. Steady swimming muscle dynamics in the leopard shark *Triakis semifasciata*. Journal of Experimental Biology. 206:1117-1126.

FELSENSTEIN, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:401-410.

FOURMANOIR, P., AND I. RIVATON. 1979. Poissons de la pente recifale externe de Nouvelle - Caledonie et des Nouvelles - Hebrides.

GARDNER, M. G., AND R. D. WARD. 2002. Taxonomic affinities within Australian and New Zealand *Mustelus* sharks (Chondrichthyes: Triakidae) inferred from allozymes, mitochondrial DNA and precaudal vertebrae counts. Copeia. 2002:356-363.

GRAY, J. E. 1851. Part I. Chondropterygii, p. 160. In: List of the specimens of fish in the collection of the British Museum. British Museum of Natural History, London.

HAMLETT, W. C., L. FISHELSON, A. BARANES, C. K. HYSELL, AND D. M. SEVER. 2002. Ultrastructural analysis of sperm storage and morphology of the oviducal gland in the Oman shark, *Iago omanensis* (Triakidae).

HEEMSTRA, P. C. 1973. A revision of the shark genus *Mustelus* (Squaliformes: Carcharhinidae), p. 187. University of Miami, Miami.

—. 1997. A review of the smooth-hound sharks (Genus *Mustelus*, Family Triakidae) of the western Atlantic Ocean, with description of two new species and a new subspecies. Bulletin of Marine Science. 60:894-928.

LAST, P. R., AND J. D. STEVENS. 1994. Sharks and rays of Australia. CSIRO, East Melbourne, Victoria.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, AND D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605-612.

SIMPFENDORFER, C. A., A. M. KITCHINGMAN, AND R. B. MCAULEY. 2002. Distribution, biology and fisheries importance of the pencil shark, *Hypogaleus hyugaensis* (Elasmobranchii: Triakidae), in the waters off south-western Australia.

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Sinauer Associates, Sunderland, Massachussetts.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN, AND D. G. HIGGINS. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl. Acids. Res. 25:4876-4882.

WHITE, E. G. 1936. A classification and phylogeny of the elasmobranch fishes. American Museum Novitates:1-16.

—. 1937. Interrelationships of the elasmobranchs with a key to the order Galea. Bulletin of the American Museum of Natural History. 74:25-138.

**Table 5.1** List of triakid and outgroup taxa sampled in this study and the genes from which DNA sequences were obtained

| Subfamily | Tribe | Genus | Species | Cyt b | ND2 | ND4 | RAG1 |
|---|---|---|---|---|---|---|---|
| Galeorhininae | Galeorhinini | *Galeorhinus* | *galeus* | X | X | X | X |
| | | *Hypogaleus* | *hyugaensis* | X | X | X | X |
| | Iagini | *Furgaleus* | *macki* | X | X | X | X |
| | | *Hemitriakis* | *japanica* | X | X | X | X |
| | | | sp. (Japan) | | | X | |
| | | *Iago* | *omanensis* | X | X | X | X |
| | | | sp. (Philippines) | X | X | X | |
| Triakinae | | *Mustelus* | *antarcticus* | | X | X | X |
| | | | *asterias* | X | X | X | X |
| | | | *californicus* | X | X | X | X |
| | | | *canis* | X | X | X | X |
| | | | *henlei* | X | X | X | X |
| | | | *manazo* | X | X | X | X |
| | | | *mosis* | X | X | X | X |

**Table 5.1** (continued)

| Subfamily | Tribe | Genus | Species | Cyt b | ND2 | ND4 | RAG1 |
|---|---|---|---|---|---|---|---|
| Triakinae | | *Mustelus* | *mustelus* | X | X | X | X |
| | | | *norrisi* | X | X | X | X |
| | | | *schmitti* | X | X | X | X |
| | | | *sinusmexicanus* | X | | X | X |
| | | *Scylliogaleus* | *quecketti* | X | X | X | X |
| | | *Triakis* | *megalopterus* | X | X | X | X |
| | | | *scyllium* | X | X | X | |
| | | | *semifasciata* | X | X | X | X |
| Outgroup taxa | | *Carcharhinus* | *acronotus* | X | X | X | X |
| | | *Prionace* | glauca | X | X | X | X |
| | | *Eusphyra* | *blochii* | X | X | X | X |
| | | *Sphyrna* | *mokarran* | X | X | X | X |
| | | *Galeocerdo* | *cuvier* | X | X | X | X |

**Table 5.1** (continued)

| | Genus | Species | Cyt b | ND2 | ND4 | RAG1 |
|---|---|---|---|---|---|---|
| Outgroup taxa | *Hemigaleus* | *microstoma* | X | X | X | X |
| | *Chaenogaleus* | *macrostoma* | X | X | X | X |
| | *Leptocharias* | *smithii* | X | X | X | X |
| | *Pseudotriakis* | *microdon* | X | X | X | X |
| | *Gollum* | *attenuatus* | X | X | X | X |
| | *Apristurus* | *macrorhynchus* | X | X | X | X |
| | *Poroderma* | *pantherinum* | X | X | X | X |

**Table 5.2** P-values of $\chi^2$ tests of base composition homogeneity for each set of gene sequences

| Gene | All sites | 3rd codon positions sites | |
|---|---|---|---|
| | | All taxa | Triakids only |
| Cytochrome *b* | 0.999 | <0.010* | 0.920 |
| NADH-2 | 0.999 | <0.050* | 0.906 |
| NADH-4 | 1.000 | 0.292 | 0.995 |
| RAG1 | 1.000 | 1.000 | 1.000 |

**Figure Captions**

**Figure 5.1** Alternative proposals of triakid relationships based on Compagno (1988b). The two arrangements differ between triakid (A) monophyly, and (B) paraphyly.

**Figure 5.2** Plots of observed transitions and transversion against uncorrected pairwise distance based on cytochrome *b* sequences at: (A) all sites, (B) 3$^{rd}$ codon position sites, and (C) 1$^{st}$ and 2$^{nd}$ position sites combined.

**Figure 5.3** Bootstrap consensus tree based on cytochrome *b* sequences with branch lengths estimated using maximum likelihood. Numbers next to nodes are percent bootstrap values.

**Figure 5.4** Plots of observed transitions and transversion against uncorrected pairwise distance based on ND2 sequences at: (A) all sites, (B) 3$^{rd}$ codon position sites, and (C) 1$^{st}$ and 2$^{nd}$ position sites combined.

**Figure 5.5** Bootstrap consensus tree based on ND2 sequences with branch lengths estimated using maximum likelihood. Numbers next to nodes are percent bootstrap values.

**Figure 5.6** Plots of observed transitions and transversion against uncorrected pairwise distance based on ND4 sequences at: (A) all sites, (B) 3$^{rd}$ codon position sites, and (C) 1$^{st}$ and 2$^{nd}$ position sites combined.

**Figure 5.7** Bootstrap consensus tree based on ND4 sequences with branch lengths estimated using maximum likelihood. Numbers next to nodes are percent bootstrap values.

**Figure 5.8** Plots of observed transitions and transversion against uncorrected pairwise distance based on RAG1 sequences at: (A) all sites, (B) 3$^{rd}$ codon position sites, and (C) 1$^{st}$ and 2$^{nd}$ position sites combined.

**Figure 5.9** Bootstrap consensus tree based on RAG1 sequences with branch lengths estimated using maximum likelihood. Numbers next to nodes are percent bootstrap values.

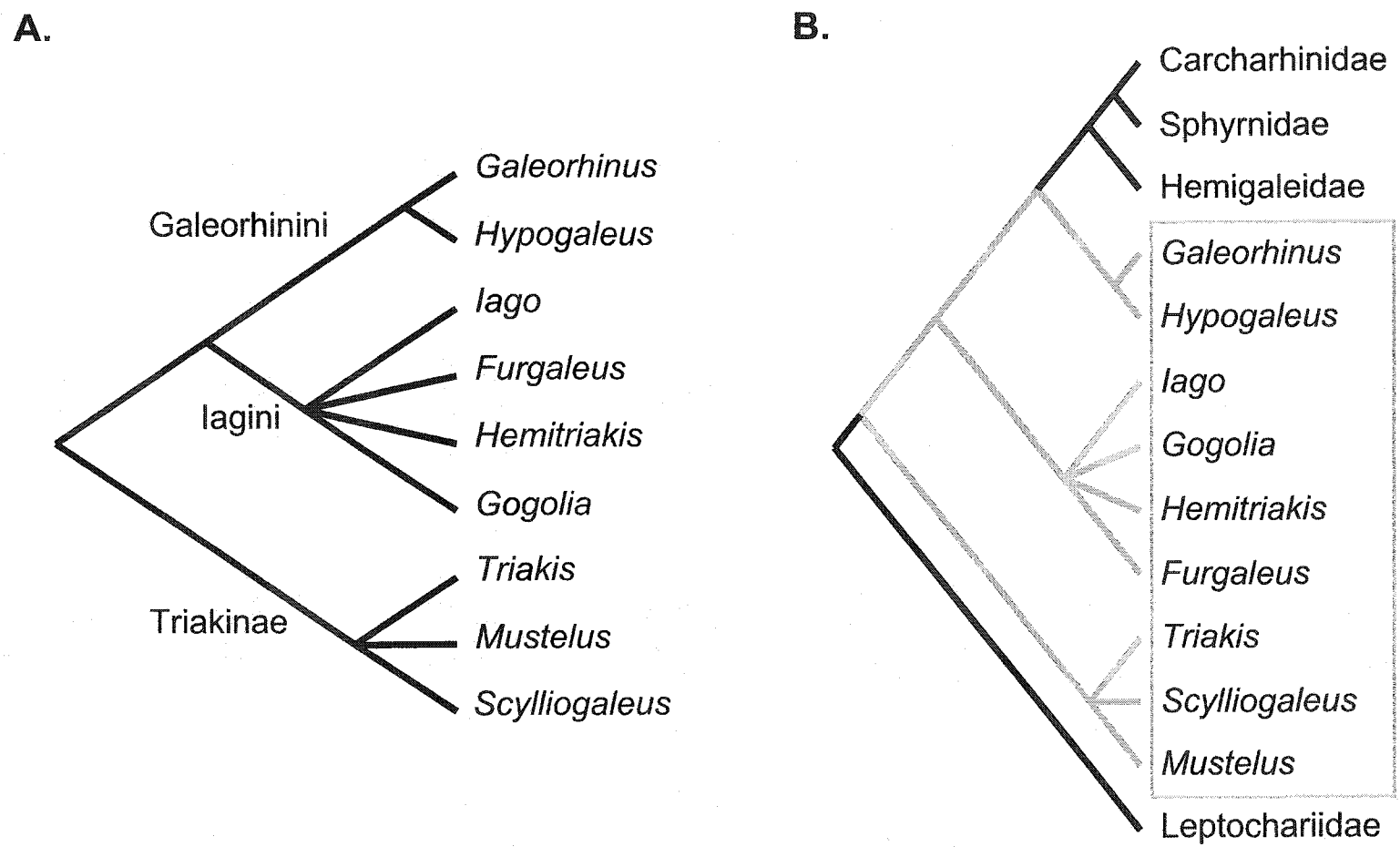
A.
Galeorhinini
Galeorhinus
Hypogaleus
Iagini
Iago
Furgaleus
Hemitriakis
Gogolia
Triakinae
Triakis
Mustelus
Scylliogaleus
B.
Carcharhinidae
Sphyrnidae
Hemigaleidae
Galeorhinus
Hypogaleus
Iago
Gogolia
Hemitriakis
Furgaleus
Triakis
Scylliogaleus
Mustelus
Leptochariidae

**Figure 5.1**

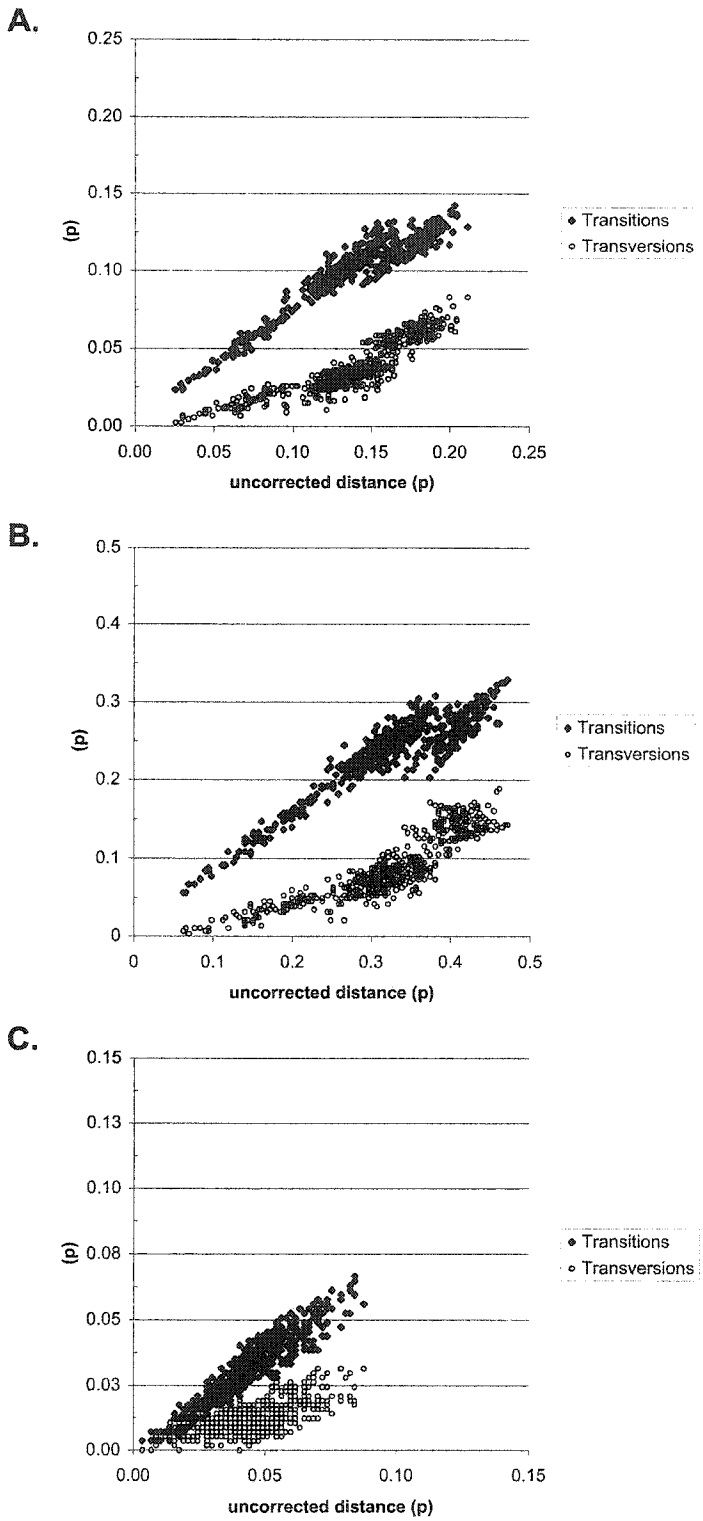
A.
0.25
0.20
0.15
0.10
0.05
0.00
(p)
0.00 0.05 0.10 0.15 0.20 0.25
uncorrected distance (p)
Transitions
Transversions
B.
0.5
0.4
0.3
0.2
0.1
0
(p)
0 0.1 0.2 0.3 0.4 0.5
uncorrected distance (p)
Transitions
Transversions
C.
0.15
0.13
0.10
0.08
0.05
0.03
0.00
(p)
0.00 0.05 0.10 0.15
uncorrected distance (p)
Transitions
Transversions

**Figure 5.2**

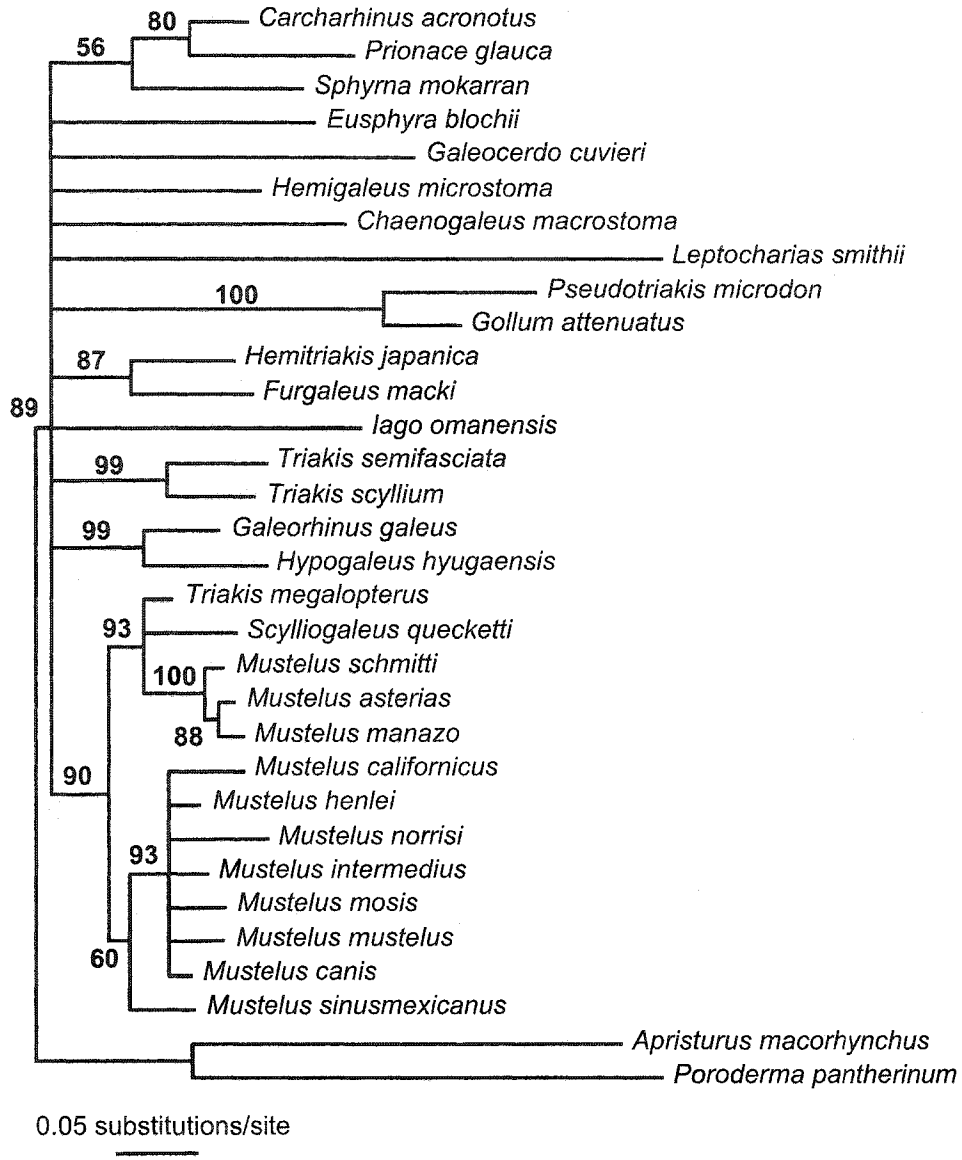
80
56
Carcharhinus acronotus
Prionace glauca
Sphyrna mokarran
Eusphyra blochii
Galeocerdo cuvieri
Hemigaleus microstoma
Chaenogaleus macrostoma
Leptocharias smithii
100
Pseudotriakis microdon
Gollum attenuatus
87
Hemitriakis japanica
Furgaleus macki
89
Iago omanensis
99
Triakis semifasciata
Triakis scyllium
99
Galeorhinus galeus
Hypogaleus hyugaensis
Triakis megalopterus
93
Scylliogaleus quecketti
100
Mustelus schmitti
Mustelus asterias
88
Mustelus manazo
90
Mustelus californicus
Mustelus henlei
Mustelus norrisi
93
Mustelus intermedius
Mustelus mosis
Mustelus mustelus
60
Mustelus canis
Mustelus sinusmexicanus
Apristurus macorhynchus
Poroderma pantherinum
0.05 substitutions/site

**Figure 5.3**

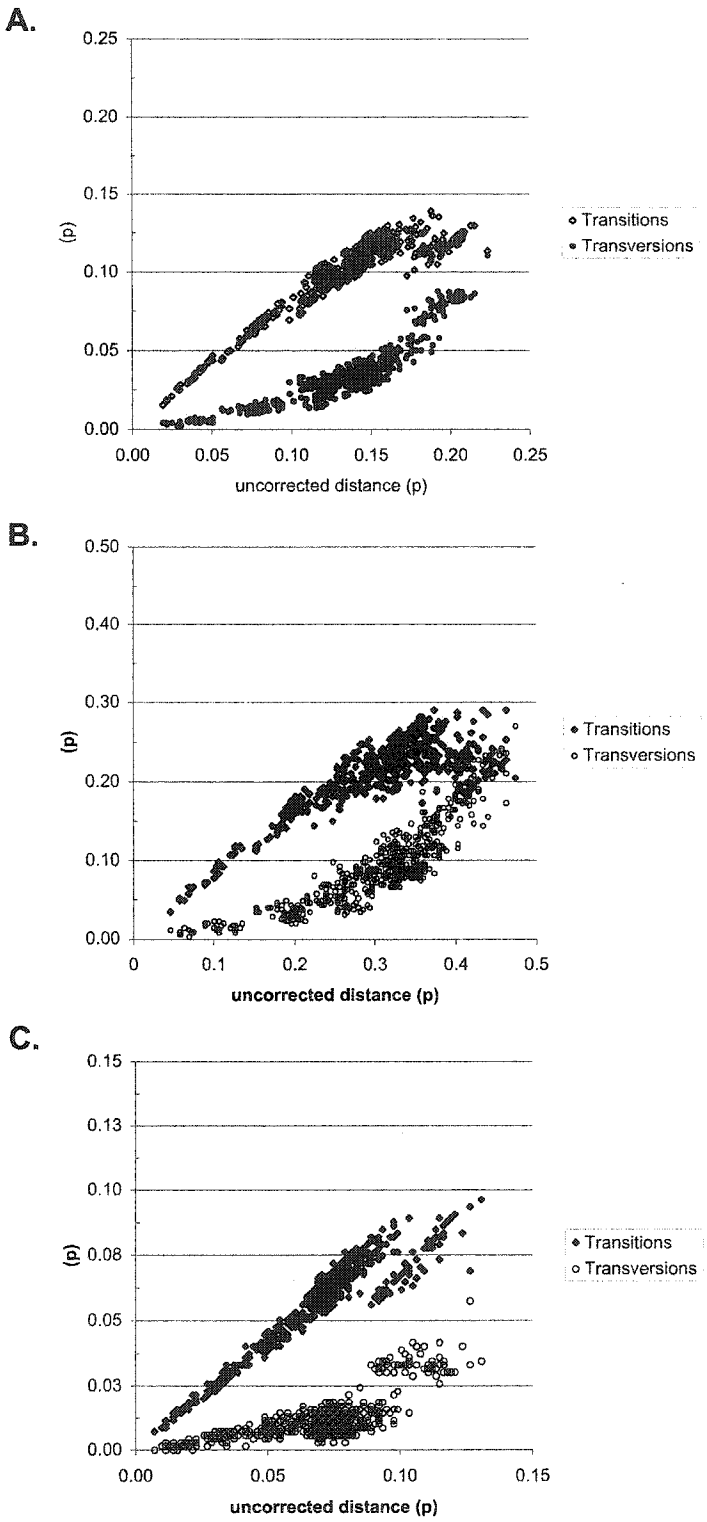
A.
0.25
0.20
0.15
0.10
0.05
0.00
(p)
0.00 0.05 0.10 0.15 0.20 0.25
uncorrected distance (p)
Transitions
Transversions
B.
0.50
0.40
0.30
0.20
0.10
0.00
(p)
0 0.1 0.2 0.3 0.4 0.5
uncorrected distance (p)
Transitions
Transversions
C.
0.15
0.13
0.10
0.08
0.05
0.03
0.00
(p)
0.00 0.05 0.10 0.15
uncorrected distance (p)
Transitions
Transversions

**Figure 5.4**

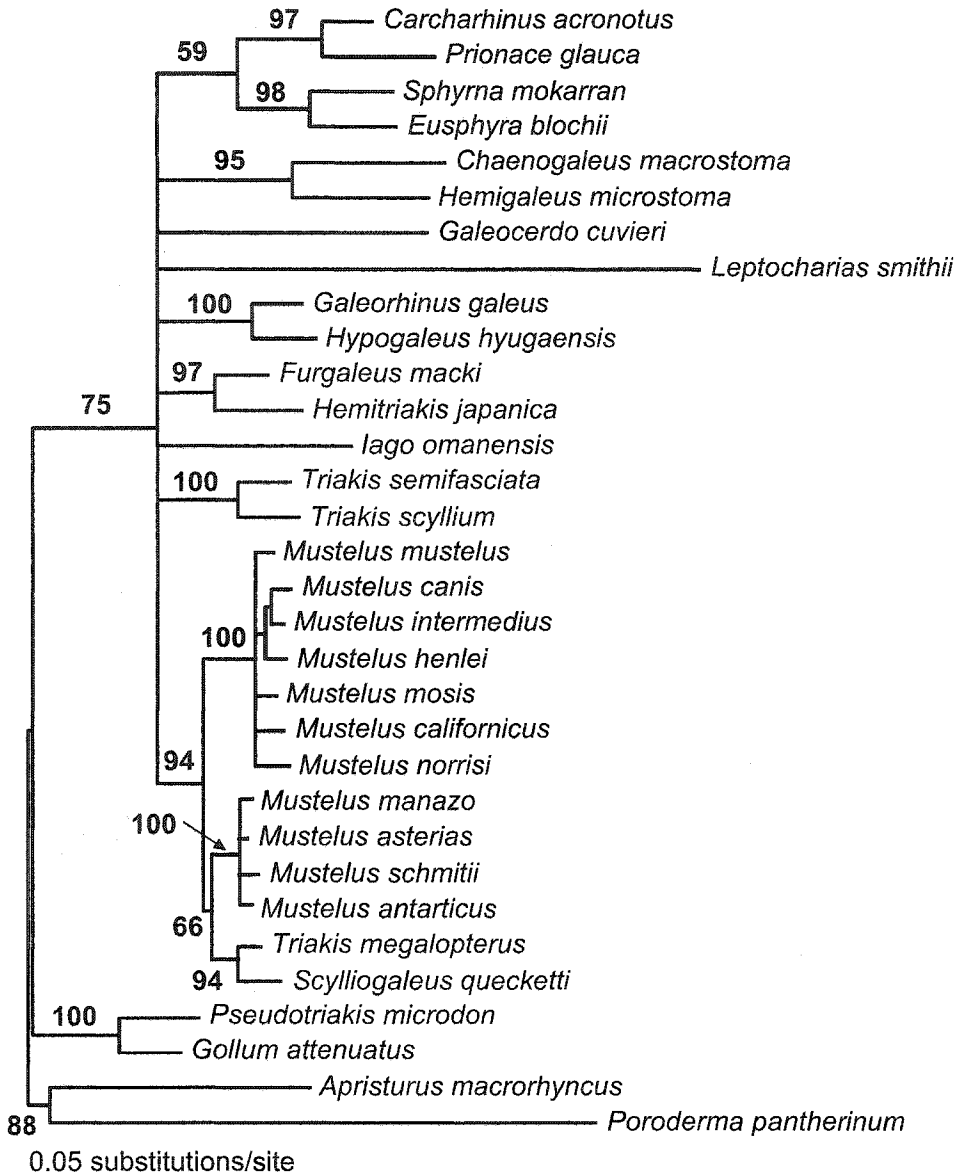
97
Carcharhinus acronotus
59
Prionace glauca
98
Sphyrna mokarran
Eusphyra blochii
95
Chaenogaleus macrostoma
Hemigaleus microstoma
Galeocerdo cuvieri
Leptocharias smithii
100
Galeorhinus galeus
Hypogaleus hyugaensis
97
Furgaleus macki
75
Hemitriakis japanica
Iago omanensis
100
Triakis semifasciata
Triakis scyllium
Mustelus mustelus
Mustelus canis
Mustelus intermedius
100
Mustelus henlei
Mustelus mosis
Mustelus californicus
94
Mustelus norrisi
Mustelus manazo
100
Mustelus asterias
Mustelus schmitii
Mustelus antarticus
66
Triakis megalopterus
94
Scylliogaleus quecketti
100
Pseudotriakis microdon
Gollum attenuatus
Apristurus macrorhyncus
88
Poroderma pantherinum
0.05 substitutions/site

**Figure 5.5**

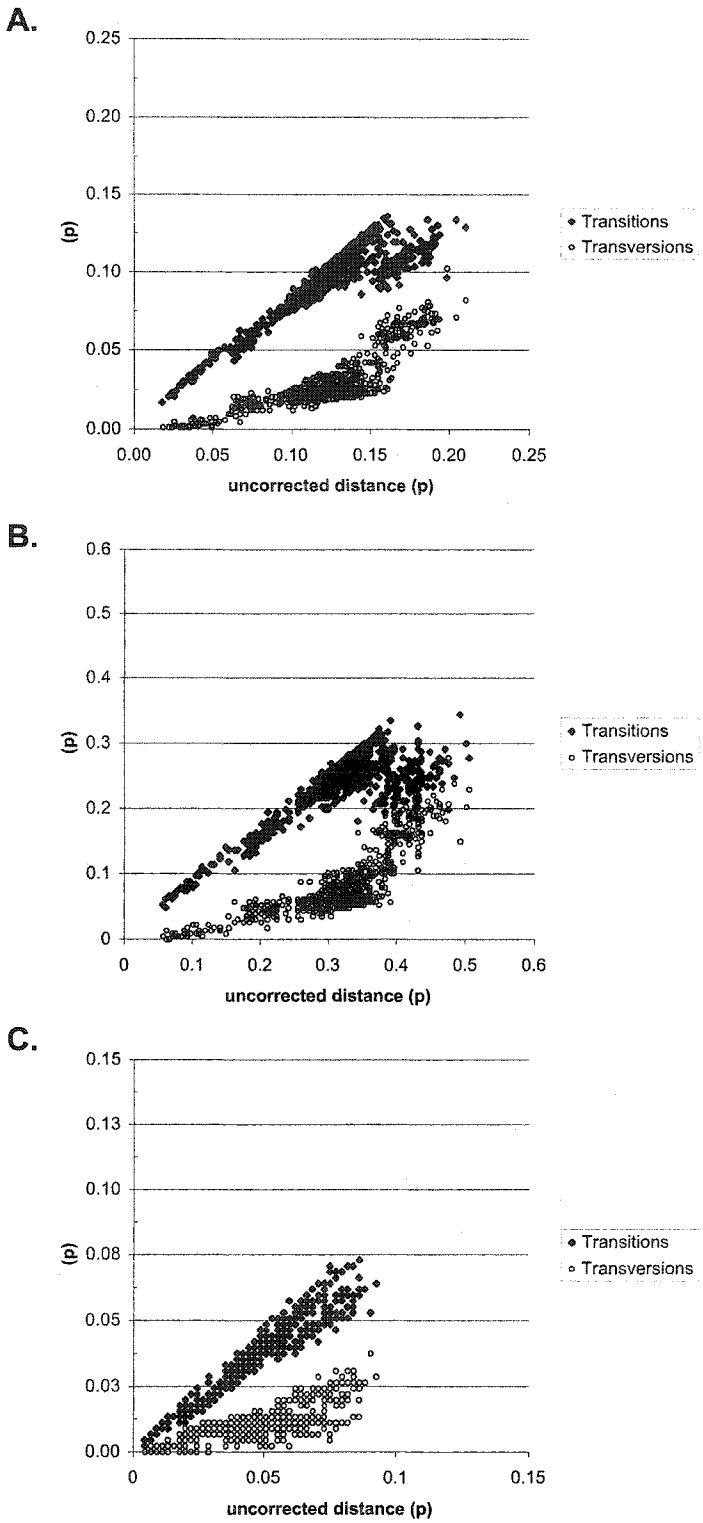
A.
0.25
0.20
0.15
0.10
0.05
0.00
(p)
0.00 0.05 0.10 0.15 0.20 0.25
uncorrected distance (p)
Transitions
Transversions
B.
0.6
0.5
0.4
0.3
0.2
0.1
0
(p)
0 0.1 0.2 0.3 0.4 0.5 0.6
uncorrected distance (p)
Transitions
Transversions
C.
0.15
0.13
0.10
0.08
0.05
0.03
0.00
(p)
0 0.05 0.1 0.15
uncorrected distance (p)
Transitions
Transversions

**Figure 5.6**

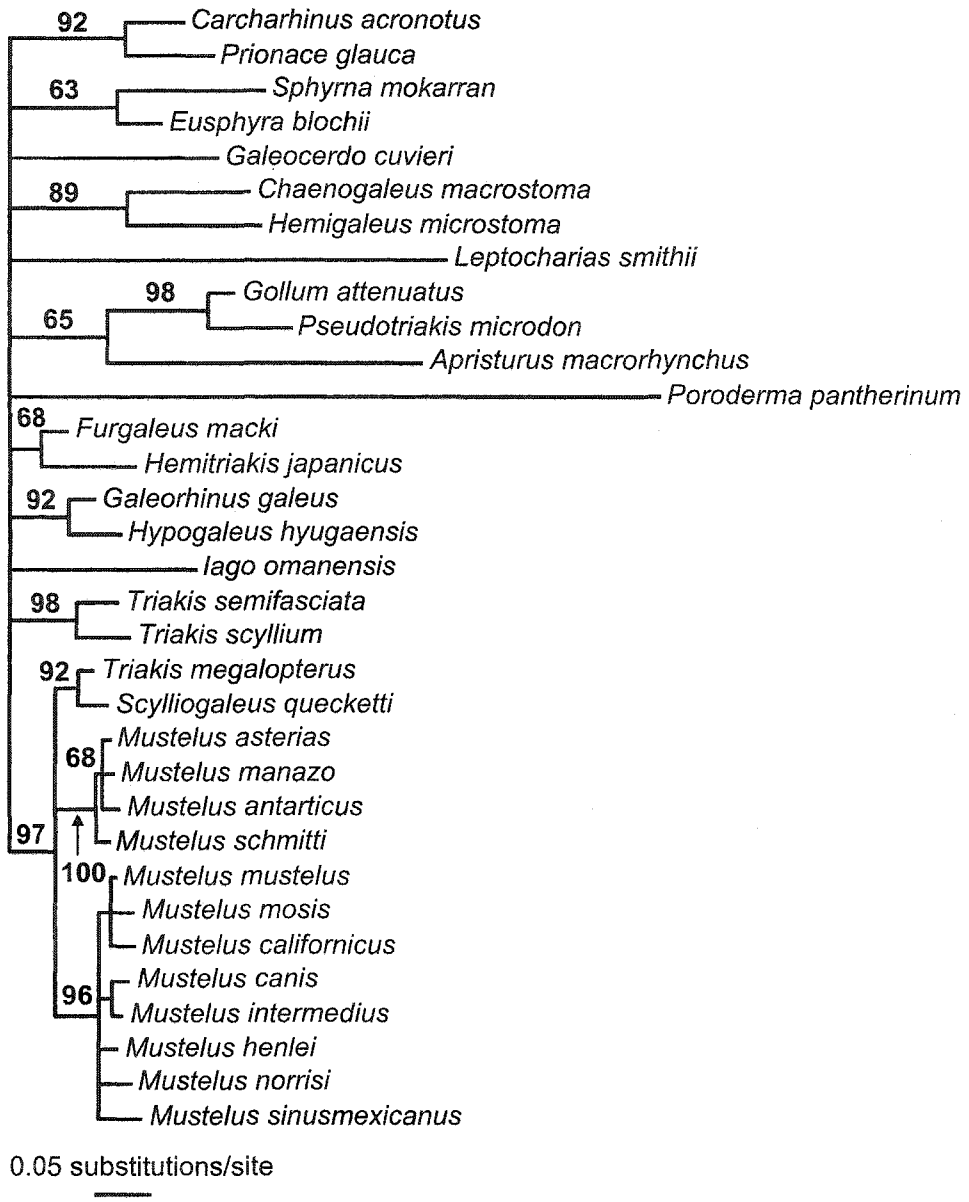
92
Carcharhinus acronotus
Prionace glauca
63
Sphyrna mokarran
Eusphyra blochii
Galeocerdo cuvieri
89
Chaenogaleus macrostoma
Hemigaleus microstoma
Leptocharias smithii
98
Gollum attenuatus
65
Pseudotriakis microdon
Apristurus macrorhynchus
Poroderma pantherinum
68
Furgaleus macki
Hemitriakis japanicus
92
Galeorhinus galeus
Hypogaleus hyugaensis
Iago omanensis
98
Triakis semifasciata
Triakis scyllium
92
Triakis megalopterus
Scylliogaleus quecketti
Mustelus asterias
68
Mustelus manazo
Mustelus antarticus
97
Mustelus schmitti
100
Mustelus mustelus
Mustelus mosis
Mustelus californicus
Mustelus canis
96
Mustelus intermedius
Mustelus henlei
Mustelus norrisi
Mustelus sinusmexicanus
0.05 substitutions/site

**Figure 5.7**

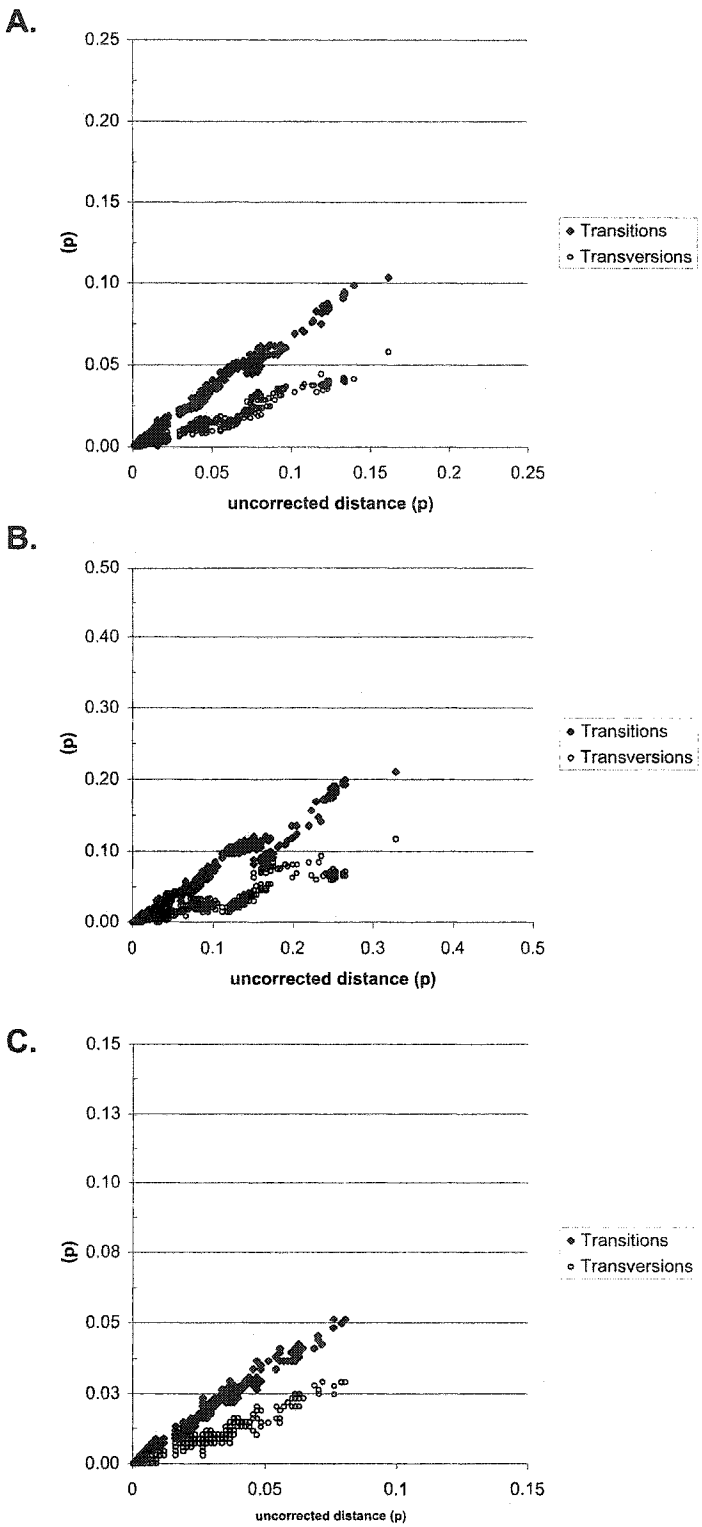
A.
0.25
0.20
0.15
0.10
0.05
0.00
(p)
0
0.05
0.1
0.15
0.2
0.25
uncorrected distance (p)
Transitions
Transversions
B.
0.50
0.40
0.30
0.20
0.10
0.00
(p)
0
0.1
0.2
0.3
0.4
0.5
uncorrected distance (p)
Transitions
Transversions
C.
0.15
0.13
0.10
0.08
0.05
0.03
0.00
(p)
0
0.05
0.1
0.15
uncorrected distance (p)
Transitions
Transversions

**Figure 5.8**

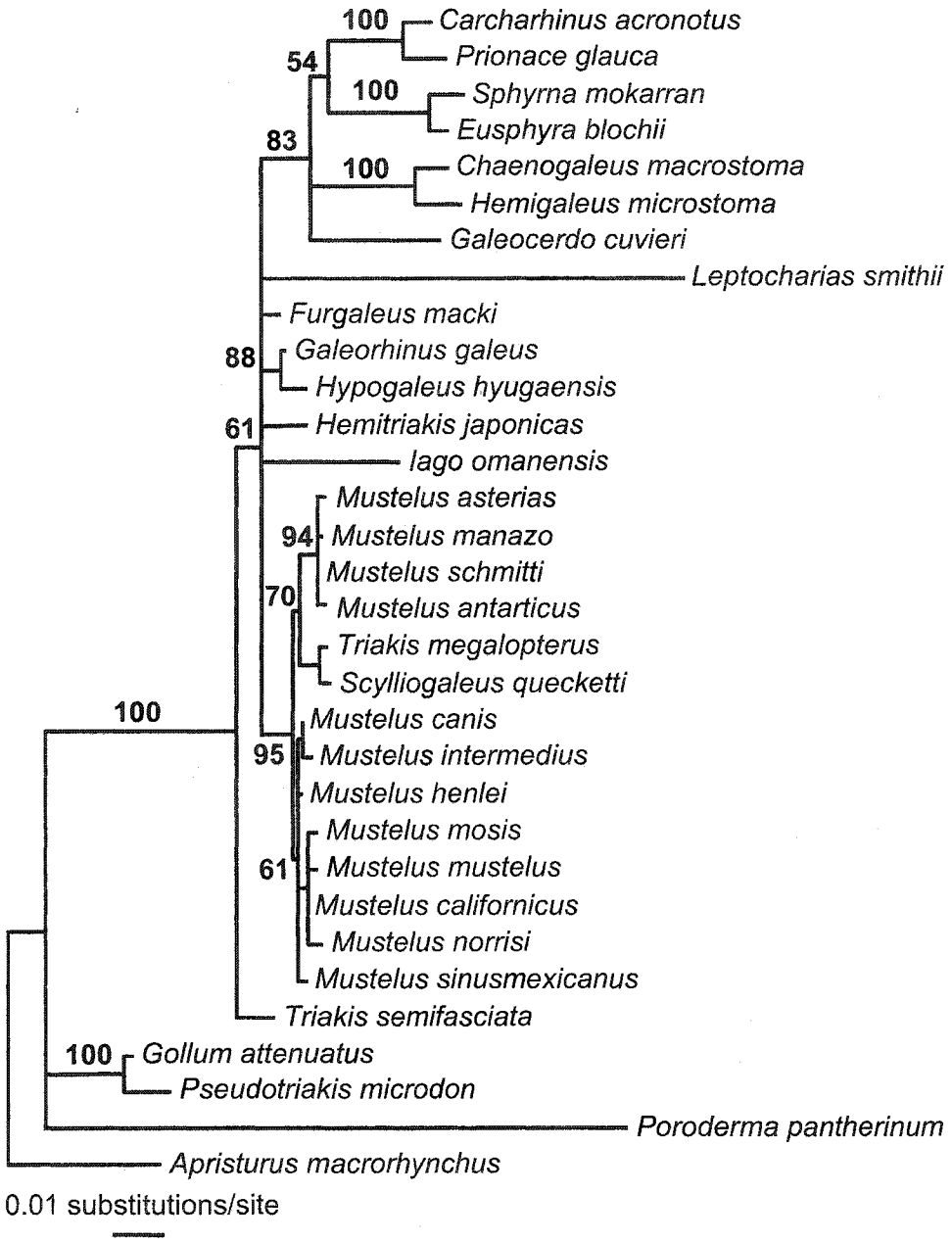
100
Carcharhinus acronotus
Prionace glauca
54
100
Sphyrna mokarran
Eusphyra blochii
83
100
Chaenogaleus macrostoma
Hemigaleus microstoma
Galeocerdo cuvieri
Leptocharias smithii
Furgaleus macki
88
Galeorhinus galeus
Hypogaleus hyugaensis
61
Hemitriakis japonicas
Iago omanensis
Mustelus asterias
94
Mustelus manazo
Mustelus schmitti
70
Mustelus antarticus
Triakis megalopterus
Scylliogaleus quecketti
100
Mustelus canis
95
Mustelus intermedius
Mustelus henlei
Mustelus mosis
61
Mustelus mustelus
Mustelus californicus
Mustelus norrisi
Mustelus sinusmexicanus
Triakis semifasciata
100
Gollum attenuatus
Pseudotriakis microdon
Poroderma pantherinum
Apristurus macrorhynchus
0.01 substitutions/site

**Figure 5.9**

# CHAPTER 6. GENERAL CONCLUSIONS

## Conclusions

Following is a summary of some of the conclusions regarding current methods and practices in phylogenetic inference as gleaned from the three case studies presented in the preceding chapters. The phylogenetic results and conclusions specific to the taxonomic groups that are the subject of those chapters are not repeated here.

Today, phylogenetic inference based on molecular sequences is an active field where new methods and flavors of methods are constantly being devised and tested against simulated and real data. Despite of, or perhaps fueled by the wealth and sophistication of approaches developed to date, many questions regarding the primacy of different methods remain open. However, what was once an acrimonious debate has turned into cordial conviviality and typically a researcher will employ a healthy variety of methods and attempt to reconcile results that are mutually contradictory or that conflict with well-established views on a particular phylogenetic question. A consequence of this methodological catholicity is that important problematic issues are commonly ignored. In the course of the studies described here, several main themes have presented themselves as recurring obstacles in the advancement of our understanding of evolutionary relationships. These themes can be broadly categorized as issues of: 1) sampling, 2) methodology and, 3) degree of confidence. Following is an outline of how these three themes are problematic for the development of hypotheses of phylogeny as commonly practiced at present.

*Sampling: by design and by nature*

Sampling presents two different types of problems to the phylogeneticist. The first type, and the most obvious, is the need to include an adequate representation of the species diversity assigned to the group under study. Defining what should be considered adequate representation for a particular taxonomic question is in principle very simple. The hypothesis being tested or question being asked dictates that every relevant taxonomic entity in that hypothesis or question be included in the test. For example, if the hypothesis being tested is the monophyly of a particular family, then every genus included in that family must be part of the test. But even if a sample meets that condition; it is also necessary that the monophyly of each of the genera assigned to the family can be assumed with some degree confidence. Although these sampling principles are easily understood, it is very common for published phylogenetic studies to make claims that the sampling design of the study do not justify. It is important to note that the consequences of inadequate sampling highlighted here are not necessarily the same as those of uneven taxonomic sample. Uneven taxonomic sampling, either by design or by necessity has a different set of negative consequences for phylogenetic inference and they are described under the second type of sampling problems.

The second type of sampling problem in the study of phylogenies is often beyond the control of the researcher. The natural patterns of speciation and extinction determine the phylogenetic distribution of the living and extinct representatives of a biological lineage on the branches of the true tree of that lineage. When the living representatives of a lineage are a poor or uneven sample of the historical diversity of that lineage even complete taxonomic sampling may be insufficient to infer an accurate phylogenetic reconstruction. This is so because the performance of currently devised methods of phylogenetic reconstruction is adversely affected by heterogeneity in the evolutionary process they intend to capture. An

uneven tree topology relating the taxa in a study can be one source of such heterogeneity. While in theory, this type of problem may only affect studies that rely on molecular data or morphological data derived exclusively from extant species, even the samples used in paleontological studies are constrained by the vagaries of fossil preservation and discovery, which can be argued are caused by systematic biases (e.g. habitat preference, location of fossil beds) rather than randomness and, therefore, may also result in an uneven sample of representative taxa from which to infer phylogeny.

It is the task of reviewers and editors to ensure that proposed phylogenetic hypothesis are logically justified by the taxonomic sampling from which they are derived; which can easily solve the issues arising from the first, of the two kinds of sampling problems outlined above. The second type of problem is much more intractable, but one significant improvement is to recognize that there are phylogenetic questions that, by the nature of the phylogenetic distribution of observable samples, may not be confidently answered (see section below). Unfortunately because most methods of phylogenetic reconstruction will find one or a few optimal trees even among a vast forest of similarly probable trees, researchers often advance those few optimal trees as working hypotheses without clearly stating the low degree of confidence that those hypotheses may warrant.

*Method selection and justification*

A survey of current literature on phylogenetic inference reveals that, with the exception of articles explicitly investigating the relative merit of different methods, most studies do not state the justification behind method selection. This is unfortunate because there are important philosophical differences behind the two main approaches to phylogenetic reconstruction (which can be termed cladistic and statistical) and these

differences have an impact on how one may gauge the strength of evidence behind a given phylogenetic proposal.

There are two very different perspectives on the nature of phylogenetic inquiry. The cladistic approach seeks an optimal tree solution to a set of observations and is little concerned with issues of confidence and the evolutionary properties of the entities being observed. On the other hand, proponents of the likelihood family of methods reflect the perspective that the problem of phylogenetic reconstruction is ultimately a statistical one, where samples of observations are collected and given an adequate such sample, an accurate inference may be produced. Therefore, under this perspective, if a method of inference can be shown to have acceptable statistical accuracy via simulated data, then it is only a question of collecting a sufficient sample of observations to obtain a good estimate of phylogeny. This perspective is responsible for the most recent advances in the field. However, while it is a useful theoretical framework, there are important ways in which the phylogeny of a particular lineage is not a statistically definable problem. A lineage has a unique history and the patterns of evolution that both, cause and are affected by that history will have unique consequences for the distribution of trait variation among the different parts of the tree. For this reason, showing that a method has statistical accuracy for one or many data sets should not give us confidence that it will conserve that property for an acceptable (e.g. the vast majority) proportion of the phylogenies that have occurred. To illustrate how this is problematic using the same statistical framework: one can conceive the true phylogenies of all lineages and sub-lineages of life that have existed on the planet as a population of phylogenies with an unknown distribution along an unknown number of characteristics (e.g. time since origin, relative diversity, patterns of extinction). If this distribution were known along all of its imaginary axis and a method could be shown to be statistically accurate for

the bulk of this distribution, then one could confidently accept the results of that method. But two problems are immediately obvious: 1) assuming that this population of phylogenies is distributed in a definable way is a strong statement that requires prior justification, and 2) a related problem, in the absence of adequate knowledge of this distribution, one cannot know with what frequency the method will produce an erroneous result. This parameter is a key component of all statistical analyses and our inability to estimate it leads to another important issue in phylogenetic inference, the issue of confidence on a hypothesis.

*Relative confidence*

Ideally, scientific hypotheses can be assigned a measure of confidence based on the weight of the existing evidence. This measure of confidence is important to those who would build on those hypotheses to investigate other related questions. In the absence of this measure, the investigation of those other question may be a misallocation of time and resources. In phylogenetic inference, the issue of confidence is problematic because of the unique nature of the question (as described above). Although measures of support, such as the bootstrap, posterior probabilities and decay indices have been devised, these only reflect the extent to which a given grouping is compatible with the evidence in the context of the assumptions implied in the method of reconstruction used. However, when different data sets offer strong support for contradictory groupings, there arises the need to weight the relative value of different types of data. For example, a data matrix of morphological characters may consist of tens or at most a few hundred such traits. Each collected and annotated through careful work by an expert over a relatively long period of time. A molecular data matrix may consist of thousands of characters, which are collected in an almost blind fashion over a short time by technicians who do not need special knowledge of

the species under study. In addition, the theoretical domain of possible evolutionary states for a morphological trait is difficult to even approximate, while that for a given nucleotide is well delineated. These are a few examples of the many ways in which different types of phylogenetic data may differ. Given such striking contrasts, how does one give a common measure of confidence to contradicting hypotheses supported by different types of evidence? This, I think, is one of the most pressing issues in phylogenetic inference.

**Future Research**

In my view, the most important avenue of development in phylogenetic inference is that which will give rise to methods of weighing different types of evidence in a common analytical framework. The case studies presented here, show that there are significant problems associated with inferring phylogeny from both morphological and molecular evidence. And, while refinements in the methods of inference will certainly be developed, most of these cannot address those fundamental issues that affect our ability to confidently accept phylogenetic hypothesis for the reasons stated above. Further, inevitable advancements in molecular biology will make available radically new kinds of relevant evidence (e.g. chromosome region homologies, repetitive element intra-genomic population relationships, variation in gene expression), which: a) will be characterized by their own unique sets of evolutionary constraints and b) by their nature exist in different relative quantities, much like gene sequences and morphological characters are available in different quantities today. This last point is important because the most abundant type of data and its constraints may disproportionately affect a combined analysis in the absence of a weighting function that can assign an accurate relative value to the different types of evidence. Therefore, one of the most important tools for the phylogeneticist of the future would provide

a means to compare bodies of different types of evidence to determine which hypothesis of relationships best fits all available evidence and assign a measure of confidence to that hypothesis. The key component of this tool will only come from increased understanding of the evolutionary properties of different types of data, which will inform the development and selection of appropriate methods of inference and, maybe more importantly, will help delineate the strengths and limitations of each type of evidence.

## ACKNOWLEDGEMENTS

I am very grateful to my family, friends and colleagues for their support, interest and encouragement throughout the development of this research, in particular, and of my work as a biologist in general. Working under the direction of Dr. Gavin ("follow from the front, lead from the rear") Naylor made graduate student life very enjoyable, stimulating and, now, very difficult to leave. He was and is always excited to discuss important questions in biology and as well as almost every thing else. Thank you Gavin. Of course, big thanks to Julie ("I don't want to play this game anymore") Ryburn who was the most entertaining lab partner I will ever have, which was both good and bad – things could and did get too entertaining sometimes. Thank you to Dean and Nicole Adams-Valenzuela biologists and hosts extraordinaire who always had on offer an oasis of warmth and friendship even in the coldest days of those long Iowa winters. Vicente ("God bless America") and Patricia brightened the lab with unrelenting optimism and the sunniest disposition that must be humanly possible. All the students that have worked in and around Gavin's lab made for an always interesting and smart group of colleagues to share the challenges and benefits of graduate student life. There are many people who helped me carry this work to completion and although not all of them are acknowledged here I hope to have and continue to express my gratitude through my actions.

On a personal level, forever I thank my family: Javier, Myriam and Emilia on whose support I know to count on unconditionally. And, of course, last but definitely not least, I thank Ellen for her love and her friendship. Having shared this part of our lives made it easy to keep things in perspective by looking forward to what is to come.

Thank you to all family, friends and colleagues; know that I am always in your debt.